Pervasive and Mobile Computing 9 (2013) 372–391

Contents lists available at SciVerse ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Application and analysis of multidimensional negative surveys in participatory sensing applications

Michael M. Groat [a,*], Benjamin Edwards [a], James Horey [b], Wenbo He [c], Stephanie Forrest [a,d]

[a] University of New Mexico, Albuquerque, NM 87131, United States
[b] Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States
[c] McGill University, Montreal, Quebec H3A 2T5, Canada
[d] Santa Fe Institute, Santa Fe, NM 87501, United States

ARTICLE INFO

Article history:
Available online 9 January 2013

Keywords:
Multidimensional data
Negative surveys
Privacy
Participatory sensing applications

ABSTRACT

Participatory sensing applications rely on individuals to share personal data to produce aggregated models and knowledge. In this setting, privacy concerns can discourage widespread adoption of new applications. We present a privacy-preserving participatory sensing scheme based on *negative surveys* for both continuous and multivariate categorical data. Without relying on encryption, our algorithms enhance the privacy of sensed data in an energy and computation efficient manner. Simulations and implementation on Android smart phones illustrate how multidimensional data can be aggregated in a useful and privacy-enhancing manner.

1. Introduction

Participatory sensing applications [1] sense, collect, analyze, and share data collected locally from a large population. They have a wide range of applications such as urban planning [2], public health [3], and vehicular transportation monitoring [4,5]. Protecting the privacy of the participants and their data is important in these applications, especially when information travels across open wireless networks. Trade-offs exist between the goal of protecting privacy and the usefulness the collected data. Energy efficiency of sensing devices is also a concern.

Existing approaches for protecting the privacy of multidimensional data [6–8] are designed for database applications, where large numbers of records from different users are available to a centralized server that summarizes statistics about the records [6,7,9,10]. However, in participatory sensing applications, individual nodes typically have access to only their own sensed values. Participants might not be willing to share sensitive information with other participants or trust a central collection server.

Our method applies negative surveys to multidimensional categorical data, where the dimensions represent different variables (e.g., temperature, time) and the categories might be symbolic values (e.g., hair color, race) or a coarse-graining of numerical data. In each dimension, the set of categories forms a proper partition of the data. Individual participants disguise their data by reporting a category chosen from the set complement of the sensed value. This operation is conducted independently for each dimension. We present algorithms that allow a base station to reconstruct the original distribution of sensed values from the disguised data [4]. This approach avoids complicated and energy intensive encryption and key management schemes.

Participatory sensing applications on wireless sensor networks (WSNs) often collect multiple observations for each sensor, for example, several different environmental values, plus time and location data. We seek to preserve the privacy of

* Corresponding author. Tel.: +1 505 277 3112; fax: +1 505 277 6927.
E-mail addresses: mgroat@cs.unm.edu (M.M. Groat), bedwards@cs.unm.edu (B. Edwards), horeyjl@ornl.gov (J. Horey), wenbohe@cs.mgill.ca (W. He), forrest@cs.unm.edu (S. Forrest).

1574-1192/$ – see front matter
doi:10.1016/j.pmcj.2012.12.004

multidimensional data such as these. For example, we describe a radiation detection scenario in Section 6.1 that estimates the distribution of radiation levels at various locations. Participants disguise both dimensions: their geographic location and their local radiation level. In some cases, data from every dimension may be considered private, and in others, values from one (non-private) dimension might reveal information about another through correlation analysis.

Our threat model assumes that the base station is *honest but curious* [11,12]. That is, we assume that it faithfully follows the network protocols but could mischievously try to collect information to use against the nodes. Additional threats come from eavesdroppers intercepting packets in transit to the base station. We do not address the problem of protecting the individual senors from adversaries.

Previous work on negative surveys required an unreasonably large number of participants to reconstruct the original data accurately [4,13]. Even small increases in the number of categories within a single dimension were problematic, and there was no attempt to address the problem of multidimensional data. This paper presents a method called dimensional adjustment, which mitigates these problems by greatly reducing the number of required data samples. Dimensional adjustment sacrifices a small amount of privacy to gain a much larger amount of utility.

When applied to continuous data, negative surveys can reconstruct probability density functions. We compare our technique to random data perturbation (RDP), which has been used in privacy-preserving data mining. We show that with sufficient samples negative surveys outperform RDP, especially when the underlying probability density functions have discontinuities.

Our protocols are implemented on Android smart phones. As a demonstration, we gridded the University of New Mexico campus into 24 different location categories and sampled the ambient sound using the phone's microphone. The campus is surrounded by streets with heavy traffic, while the interior regions are relatively quiet. The goal of the experiment was to distinguish these noisier boundary locations from quieter locations on campus.

The main contributions of this paper include the following.[1] (1) Algorithms and implementations for protecting the privacy of multidimensional sensor data, in particular, focusing on participatory sensing applications that report to a base station physical location together with sensor values. (2) Algorithms that guarantee efficient and accurate reconstruction of disguised data at the base station. (3) Simulations and a prototype implementation on Android phones that demonstrate the practicality of the method under various application scenarios. (4) The dimensional adjustment technique, which reduces (from earlier work) the number of participants required to maintain a given level of utility. This technique can also be used for single-dimensional data to improve accuracy.

Roadmap: In the remainder of the paper, we first review related work, giving background information on negative surveys and randomized response techniques (Section 2). Our protocols are presented in Section 3, and Section 4 describes the privacy and utility metrics used in the analysis. Dimensional adjustment is introduced and analyzed in Section 5. Section 6 introduces a cell phone radiation detection simulation and a simulation of MDNSs on continuous data, which is compared to random data perturbation. This section also describes the implementation of MDNS on Android smart phones. We discuss the benefits and limitations of our algorithms in Section 7, speculating about how to improve the performance and security of the simulations and implementation in future work. Section 8 concludes the paper.

2. Related work

Privacy-preserving algorithms have been developed for data mining [14–17], data aggregation [4,18–22], and other applications [23,24]. There are four main classes of solutions: perturbation, k-anonymity, secure multi-party computation, and homomorphic encryption. We review these briefly, focusing on random response techniques, including a specific instance known as "negative surveys", which serve as important background for the new algorithms and results presented in this paper.

In data mining, data values are typically hidden by perturbing individual data or query results [14–16]. To obtain accurate results, these methods typically assume that the distribution of data/noise is known ahead of time. However, as shown by Kargupta et al. [14] and Huang et al. [16], certain types of data perturbation might not preserve privacy well.

The k-anonymization technique [6,7,9,10] makes a data value from a participant indistinguishable from $k-1$ other items. It was originally designed for privacy-preserving data mining, but in participatory sensing applications individual participants can sense and share their own data. Thus, there is limited potential to mix individual participants' data as required for k-anonymity.

Secure multi-party computation (SMC) [25–27] methods specify a joint computation among a set of involved peers. This is problematic in a participatory sensing setting, because of high communication or computation overhead when the participant population is large.

[1] An earlier version of the paper appeared in PERCOM'12. We have expanded the paper to include the following material. (1) Implementation of multidimensional negative surveys (MDNSs) on smart phones. This includes comparing the energy use of the node protocol with and without encryption on a phone, and conducting an experiment to map noise levels on the University of New Mexico campus. (2) A comparison of negative surveys that operate on continuous data to random data perturbation. (3) A section detailing how the original data distribution affects our privacy and utility metrics. (4) A section using the Kronecker technique to confirm the correctness of our metrics in terms of the variance and covariance of MDNS. (5) Increased detail about the dimensional adjustment technique, including an explanation of how errors can be magnified with negative surveys, and a proof that dimensional adjustment always improves utility. Finally, we have expanded the literature review of earlier.