Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Spatial counts under differential privacy mechanism on changing spatial scales

Jun Jiang [a], Bowei Xi [b,*], Murat Kantarcioglu [c]

[a] School of Engineering and Computer Science, Washington State University Vancouver, USA
[b] Department of Statistics, Purdue University, USA
[c] Department of Computer Science, University of Texas at Dallas, USA

ARTICLE INFO

Article history:
Available online

Keywords:
Spatial counts
Differential privacy
Poisson-Binomial distribution
Data generation
Transformation invariance

ABSTRACT

With a spatial statistical database covering a large region, how to publish differential privacy protected information is a challenge. In previous works, information was published using large fixed spatial cells. In this paper, we develop novel flexible methods to publish the spatial information, which allows the users to freely move around the large region, zoom in and zoom out at arbitrary locations, and obtain information over spatial areas both large and small. We develop two methods to publish the spatial information protected under differential privacy. First the region is divided into the smallest spatial cells, where each cell does not observe an event happening more than once. Given repeated measurements, such as multiple day data, the noise added Bernoulli probabilities are computed for all the smallest spatial cells. For larger spatial cells of high interests to users, the noisy Bernoulli probabilities are combined into noisy Poisson-Binomial distributions which also satisfy differential privacy requirement. We use the New York Taxi data in the experiments to demonstrate how our methods work. We show that both of our methods are accurate, while the noisy count probabilities directly obtained from fixed large spatial cells often generate the spatial counts much smaller than the true values.

1. Introduction

We consider a spatial statistical database where the spatial counts needs to be published under differential privacy protection. For example, with New York Taxi data (New York City Taxi Trip Data, 2010–2013), we are interested in the numbers of taxi pick-ups in different spatial regions/locations. How to publish the information about spatial counts over a large region under differential privacy is not as straight forward as publish differentially private one dimensional counts, where there has been some well studied work, such as differentially private histograms. Since spatial data normally covers a large region, the past work (Mir et al., 2013; Wang et al., 2016) used large fixed spatial cells and published differentially private counts in those fixed large spatial cells.

In this paper we develop two methods to publish information from a spatial database under differential privacy protection. Our methods do not use a fixed large spatial cell size as in the previous works. Instead we propose more flexible approach. Just as users can zoom in and zoom out on Google maps to have a better view of a certain spatial region, the published information using our approach allows users to view spatial information over spatial regions both small and large. Users also can freely move around the map without blackout spots. On the other hand, using fixed size large spatial cells

* Corresponding author.
E-mail address: xbw@purdue.edu (B. Xi).

https://doi.org/10.1016/j.cose.2017.11.018

means users cannot obtain any information for areas smaller than the pre-determined cell size. An area which is larger than the fixed cell size, but sits across two or more cells, also becomes black-out spots where users cannot obtain any information.

We develop two methods to publish the probabilities of spatial counts, both satisfying $\epsilon-$ differential privacy. For the first method we divide a large region into very small spatial cells. The size of the smallest spatial cells needs to be chosen carefully. With the New York Taxi data, we use 5 meter by 5 meter spatial cells. In all these small cells, the probability of observing an event happening more than once is negligible, considered as zero. Meanwhile the size of these small cells needs to be as large as possible, so the probability of observing one event is not too close to zero. This would also reduce the amount of information (i.e., the number of Bernoulli probabilities) to be released. Given repeated measurements (e.g., multiple day data or multi-hour data), the Bernoulli probabilities of observing an event are computed with added Laplace noises for all the small cells covering the region. The noise added Bernoulli probabilities are then published. In the big data era, storing a large number of Bernoulli probabilities is not a difficult task.

For a spatial cell covering more than one smallest cell, we combine the noisy Bernoulli probabilities into a Poisson-Binomial distribution. Note that the actual numbers of cells larger than the smallest cells on a map is exceedingly large. We recommend that the noisy Poisson-Binomial distributions are computed and stored only for larger spatial cells of high interests to users. We show the noisy Poisson-Binomial distributions also satisfy $\epsilon-$ differential privacy. Then through experiments with the New York Taxi data, we compare our methods, the Bernoulli method and the Poisson-Binomial method, with the noisy count probabilities over different spatial cell sizes and using different ϵ values. Both our methods are more accurate than the noisy count probabilities.

The paper is organized as follows. In Section 1.1 we discuss the related work. Section 2 introduces the differential privacy mechanism. In Section 3 we discuss the Bernoulli method and the Poisson-Binomial method. In Section 4 we conduct experiments using the New York Taxi data to compare our methods with the noisy count probabilities. Section 5 concludes this paper.

1.1. *Related work*

One approach to release differentially private count distribution is to publish a differentially private histogram. A histogram combines the counts into several bins. The number of bins and the bin size are two important factors for a differentially private histogram. Dwork et al. (2006) first introduced the concept of differentially private histogram, and provided a relatively straight forward approach. Machanavajjhala et al. (2008) considered differentially private histogram under a Bayesian framework. They had Dirichlet prior and posterior for the bin probabilities. They established a constraint for the posterior to ensure the perturbed histogram satisfies differential privacy requirement. Wasserman and Zhou (2010) studied several differentially private histograms and analyzed their convergence rate under both L_2 distance and Kolmogorov–Smirnov distance. Blum et al. (2013) proposed to have such bin sizes that the sum of counts in the bins are nearly the same. Hay et al. (2010) proposed an approach to reduce the variance of the noise for the query responses. Xiao et al. (2011) developed a wavelet method to handle multi-dimensional data with a low noise variance upper bound. Xu et al. (2013) introduced two algorithms to improve the accuracy of differentially private histograms.

However histogram is a less accurate method to publish a count distribution. Directly adding Laplace noise to basic queries, such as count and mean, appeared early in differential privacy literature (Dwork, 2008). Earliest work (Dwork et al., 2006) also considered adding Gaussian noise, Poisson noise to such query responses. In this paper we compare our methods with the noisy count probabilities which are published directly without being grouped into histogram. We show through experiments that a noisy Poisson-Binomial distribution constructed using the noisy Bernoulli probabilities is more accurate than the noisy count probabilities.

Wang et al. (2016) developed a mechanism to release spatial-temporal data under differential privacy. They started with large regions, such as 80 meters by 110 meters for Taxi Trajectory data. Then the regions with small statistics values were further grouped together. Instead, our work shows directly publishing statistics of the smallest spatial cells achieves very accurate results. It is also a much more flexible approach to allow the viewers to see the responses over spatial regions of any size and in arbitrary locations. Mir et al. (2013) developed a mechanism to publish differentially private information from cell phone call detail records. The spatial cells used were 0.01 degree of longitude by 0.01 degree of latitude or larger, roughly 1100 meters by 800 meters or larger. They generated synthetic data using their approach and compared with real data. The differences were on a scale of 0.17–2.2 miles in distance by using very large spatial cells.

2. Differential privacy mechanism

Differential privacy mechanism (Dwork, 2008; Dwork and Smith, 2010; Dwork et al., 2006) releases aggregate information from a statistical database, ensuring an individual participant's information cannot be discovered while entering or leaving the database. A statistical database can be queried in both an interactive setting or an non-interactive setting. The definition follows a rigorous mathematical framework. Differential privacy is achieved by injecting noise to the response of a query, while making the distributions of the responses over two databases differing by one element nearly identical. Either Laplace mechanism or exponential mechanism can be applied to achieve $\epsilon-$ differential privacy. Exponential mechanism applies to the non-numeric queries, while Laplace mechanism applies to the query functions with numerical value outputs. According to (Dwork, 2008; Dwork and Smith, 2010; Dwork et al., 2006), $\epsilon-$ differential privacy is defined as follows. Let K be a randomized function, and the difference between two databases D and D' is at most one element. If $\forall\ B \in range(K)$,

$$\frac{Pr(K(D) \in B)}{Pr(K(D') \in B)} \le e^{\epsilon},$$

then K satisfies $\epsilon-$ differential privacy. Differential privacy has several useful properties. In particular, it has transformation invariance (Kifer and Lin, 2010), defined as follows.