Available online at www.sciencedirect.com

ScienceDirect

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

ELSEVIER

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
0167-4048/© 2017 Elsevier Ltd. All rights reserved.

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), ϵ – 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.

Definition 1. (Transformation Invariance (Kifer and Lin, 2010)). *Let K be a differential privacy mechanism, and let A be a randomized algorithm whose input space is the output space of K. If the randomness of one algorithm is independent of the other, K′ = A o K also satisfies* $\epsilon-$ *differential privacy.*

A key concept in differential privacy is the sensitivity of a query function *f*. *f*'s sensitivity is defined as the maximum change in function *f* value over two databases differing by at most one element. Hence sensitivity is only related to the function *f* itself. Sensitivities of many query functions are well established. For example, count has sensitivity 1, and mean has sensitivity 1/*N*, *N* being the number of objects in the database.

Laplace mechanism adds a Laplace noise to the output of a query function *f*. The noise added response to the query function *f* is

$$f(D)+Lap\left(\frac{\text{sensitivity}}{\epsilon}\right).$$

ϵ controls the strength of the privacy guarantee. It is a predetermined important parameter. Dwork and Smith (2010) suggested that ϵ could be as small as 0.01, 0.1, or in some other cases, ln(2) and ln(3). Hsu et al. (2014) mentioned ϵ was chosen as small as 0.01 (Sarwate et al., 2009), as large as 7 (Machanavajjhala et al., 2011) in the literature. In this paper, we evaluate our approach using ϵ values from 0.05 to 4.

3. Materials and methods: spatial counts on changing spatial scales

When a map of synthetic spatial counts is constructed based on the responses to the queries to a statistical spatial database, the map needs to allow the map viewers to freely zoom in and out and view the resulting counts. The map should also allow the viewers to freely move around the map and view the counts on different spatial scales. Given the availability of repeated measurements, and given the goal of constructing an interactive map on varying spatial scales, one method to query the spatial database is first to focus on small spatial cells. In these small spatial cells, the probabilities of observing an event happening two or more times are negligible. The queries then is to seek the Bernoulli probabilities of an event in these small spatial cells.

With the noise added Bernoulli probabilities, if certain larger spatial cells are of interest to many map viewers, the noise added Bernoulli probabilities can be combined into Poisson-Binomial distributions over larger cells. An alternative is to directly query and obtain the noise added count probabilities over certain larger cells. The probabilities can be used to generate synthetic spatial counts, which are organized into a map for the viewers. Below we describe our Bernoulli method, our Poisson-Binomial method, and the noisy count probabilities.

3.1. *Bernoulli probabilities for smallest spatial cells*

In a spatial database, assume *T* repeated measurements of the spatial counts are available, for example, *T* days of the spatial data captured and stored. We carefully choose a smallest spatial cell size C^{min}, which is *A* meters by *A* meters. In such smallest spatial cells, the probability of observing an event happening more than once is negligible (i.e., *Pr*(*happening more than once*) = 0). Meanwhile the size of the smallest spatial cells need to be large enough, such that probabilities of observing an event are not too small. Often the scales of an interactive map is public knowledge, known to both the regular users and the adversaries. Hence C^{min} is also known to everyone.

Let C_{ij}^{min} be the (*i*, *j*) – th smallest spatial cell on a map. Let $m_{ij}(1)$ be the count of observing an event in the cell, and $m_{ij}(0)$ be the count of not observing an event.

$$m_{ij}(0)+m_{ij}(1)=T.$$

Count has sensitivity 1. Under the Laplace mechanism, a noise δ_{ij}^{min} can be directly added to the count $m_{ij}(1)$, with $\delta_{ij}^{min}\sim Lap\left(\frac{1}{\epsilon_{min}}\right)$. The noise added count for cell C_{ij}^{min} is

$$m_{ij}^{d}(1)=m_{ij}(1)+Lap\left(\frac{1}{\epsilon_{min}}\right).$$

Equivalently, the probability of observing an event is the mean of *T* Bernoulli (0/1) events.

$$\hat{p}_{ij}=\frac{m_{ij}(1)}{T}.$$

Mean has sensitivity 1/*T*. Hence the noise added probability of observing an event is

$$p_{ij}^{d}=\frac{m_{ij}(1)}{T}+Lap\left(\frac{1}{T\epsilon_{min}}\right)=\frac{m_{ij}(1)+Lap\left(\frac{1}{\epsilon_{min}}\right)}{T}. \tag{1}$$

The probability of event not happening in the cell is $1-p_{ij}^{d}$. The map area is divided into the smallest spatial cells. Queries can be sent to a spatial database to obtain the above noisy Bernoulli probabilities for all these smallest spatial cells. There are special cases, where an event cannot happen in a particular cell, we have $m_{ij}(0)=T$. For example, a taxi pick-up cannot happen inside a building. Then both $p_{ij}^{d}=0$ and $m_{ij}^{d}(1)=0$ for the cell. Obviously the noisy Bernoulli probabilities satisfy $\epsilon-$ differential privacy.

3.2. *Poisson-Binomial distribution for larger spatial cell*

For a larger spatial cell, the noisy Bernoulli probabilities of the smallest cells inside the larger cell can be organized into a distribution which can be stored and used later.

As an example, consider the case when the smallest spatial cells C^{min} are 5 meters by 5 meters, and a large spatial cell C^{L} which is 50 meters by 50 meters. C^{L} contains $M=100$ smallest cells C^{min}. Let X_i be Bernoulli random variable with probability of observing an event equal to p_i^{d} (Equation 1), with $i=1,\ldots,M$. Let *Y* be the sum of independent Bernoulli random variables X_is.

$$Y = \sum_{1}^{M} X_i.$$

Whether an event happens or not in a smallest spatial cell follows the distribution of X_i. The distribution of the number of events observed in the larger cell C^L follows the distribution of Y under differential privacy mechanism. If the noisy Bernoulli probabilities were constant, $p_1^d = \ldots = p_M^d = p$, Y would be a standard binomial random variable. Because the noisy Bernoulli probabilities are different for different cells, Y, the sum of independent but not identically distributed Bernoulli random variables, follows a Poisson-Binomial distribution (Hong, 2013).

Poisson approximation or normal approximation can be used to compute the cumulative distribution function (cdf) of the Poisson-Binomial random variable Y (Hodges and Le Cam, 1960; Le Cam, 1960). In this paper, we use the exact and closed form formula for Y derived in Fernndez and Williams (2010) and Hong (2013). Note Y takes values from 0 to M. Let $\mathbf{i} = \sqrt{-1}$. Let $\omega = 2\pi/(M+1)$. Let

$$\gamma_j^d = \prod_{l=1}^{M} \left[1 - p_l^d + p_l^d \times exp(\mathbf{i}\omega j)\right],\ j = 0, \ldots, M.$$

The cdf of Y is as follows.

$$Pr(Y \le y) = \frac{1}{M+1} \sum_{j=0}^{M} \frac{[1 - exp(-\mathbf{i}\omega j(y+1))]\gamma_j^d}{1 - exp(-\mathbf{i}\omega j)}$$

We implement the fast Fourier transform algorithm developed in Hong (2013) to compute the cdf of Y for a larger cell. And subsequently we use the noisy cdf of Y to generate synthetic counts in a larger cell following a noisy Poisson-Binomial distribution.

Remark. Due to the transformation invariance property (Definition 1) of differential privacy, the cdf of the Poisson-Binomial distribution also satisfies ϵ– differential privacy.

3.3. *Count probability for larger spatial cell*

Assume a larger cell C^L is the size of M smallest spatial cells C^{min}. Assume up to b events is observed in C^L in a single measurement. The counts in cell C^L add up to T.

$$m_L(0) + m_L(1) + \ldots + m_L(b) = T.$$

Count has sensitivity 1. The noise added counts are

$$m_L^d(k) = m_L(k) + Lap\left(\frac{1}{\epsilon_L}\right). \tag{2}$$

Consequently the noisy probability of observing an event k times in cell C^L is

$$p_L^d(k) = \frac{m_L^d(k)}{\sum_0^b (m_L^d(k))} = \frac{m_L(k) + Lap\left(\frac{1}{\epsilon_L}\right)}{\sum_0^b \left(m_L(k) + Lap\left(\frac{1}{\epsilon_L}\right)\right)}. \tag{3}$$

Remark. Due to the transformation invariance property (Definition 1) of differential privacy, the noisy count probabilities $p_L^d(k)$s also satisfy ϵ– differential privacy.

Matching Noises on Different Spatial Scales To ensure the overall noise level in an area stay reasonably constant for different spatial cell sizes, we match the variances of a larger spatial cell C^L with the total variances of the M smallest spatial cells contained in C^L. A Laplace noise has variance $1/\epsilon^2$. We set

$$\frac{1}{\epsilon_L^2} = \frac{M}{\epsilon_{min}^2}.$$

Equivalently we set the ϵ_{min} for the smallest spatial cells C^{min} as $\sqrt{M}$ times the ϵ_L of a larger spatial cell.

$$\epsilon_{min} = \sqrt{M}\epsilon_L.$$

4. Results and discussion: New York taxi data

We use two months of 2013 New York Taxi data (Donovan and Work, 2015), August and September, to compare our methods and the noisy count probabilities. Taxi data is publicly available and can be downloaded from (New York City Taxi Trip Data, 2010–2013). New York taxis are equipped with GPS device, which periodically send the GPS updates. In this dataset, taxi trips with passengers in the cars are recorded. A record of a taxi trip has eight variables, 1) pick-up time, accurate up to a second; 2) pick-up longitude; 3) pick-up latitude; 4) drop-off time, also accurate up to a second; 5) drop-off longitude; 6) drop-off latitude; 7) duration of the trip; 8) distance of the trip.

In this paper we focus on the taxi pick-up events, which are used to show the heavy traffic area in New York City. In our experiment, we focus on the area surrounding Grand Central, a busy area. We first convert the longitude and latitude of a taxi pick-up event to (x, y) coordinates measured in meters, which facilitates the construction and computation of spatial cells and the counts of events in spatial cells. We set the origin at Grand Central. The longitude and latitude of the origin (0, 0) is (–73.9765, 40.7528). Then we compute the lengths of one degree of longitude and one degree of latitude measured in meters at the origin, following the formulas of great-circle distance (Great-circle distance ; Latitude). Given latitude $\phi = 40.7528$, we have

$$d_{lat} = 111132.954 - 559.822 \times \cos\left(\frac{2\phi}{180\pi}\right) + 1.175 \times \cos\left(\frac{4\phi}{180\pi}\right)$$

$$d_{Lon} = \frac{\pi \times 6378137 \times \cos\left(\frac{\phi}{180\pi}\right)}{180\sqrt{1 - 0.00669437999014 \times \sin^2\left(\frac{\phi}{180\pi}\right)}}$$

At Grand Central, one degree of longitude measures 84448.739463 meters, and one degree of latitude measures 111049.137430 meters.

$$d_{lat} = 111049.137430, \qquad d_{Lon} = 84448.739463.$$

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018

Hence a taxi pick-up (*x*, *y*) coordinates are computed as follows.

$$x = (pickup.lon + 73.9765) \times d_{Lon},$$

$$y = (pickup.lat - 40.7528) \times d_{lat}.$$

4.1. *Experiments*

We use the August 2013 New York Taxi data to compute the noisy counts and the noisy Bernoulli probabilities. We create repeated measurements ourselves. We take the 2pm–4:30pm period from Monday to Thursday in August, a relatively stable period of the day, and break the taxi trip records into 5 minute intervals. Thus we obtain 660 repeated measurements ($T = 660$) of the spatial events, i.e., taxi pick-ups. We choose the smallest spatial cell C^{min} as 5 meters by 5 meters. For nearly all the smallest spatial cell, we observe at most one taxi pick-up in each 5 minute interval (i.e., one measurement). The size of C^{min} is also large enough so the Bernoulli probability is not too close to zero. Given a ϵ_{min}, we then compute the noisy Bernoulli probabilities p_{ij}^{d} (Equation 1) for all the smallest spatial cells surrounding Grand Central. We assume the time and

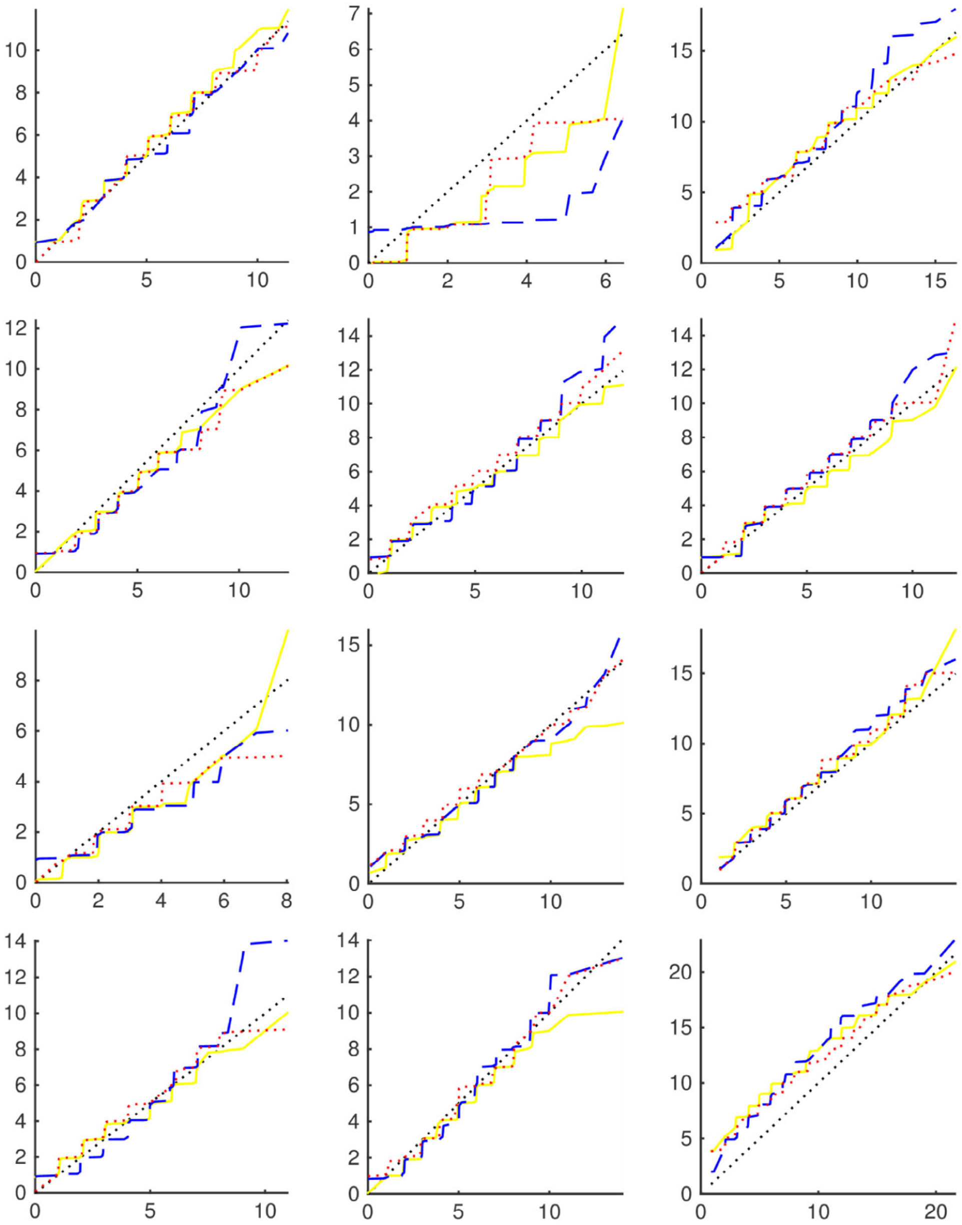

Fig. 1 – Comparing synthetic counts from three methods with the real counts in September. $\epsilon_{min} = 4$. Spatial cell size 200 meters by 200 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 200m*200m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

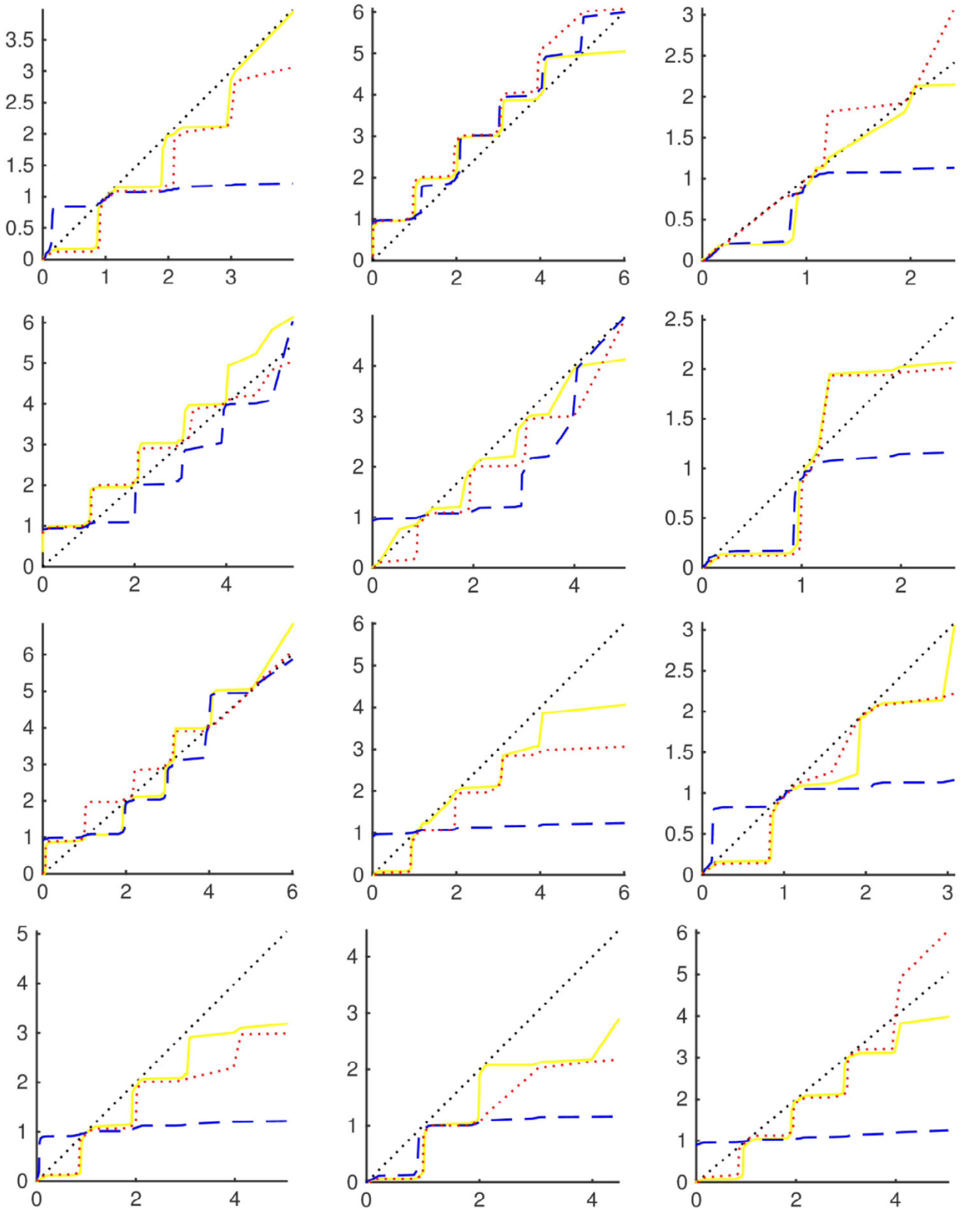

Fig. 2 – Comparing synthetic counts from three methods with the real counts in September. $\epsilon_{min} = 4$. Spatial cell size 100 meters by 100 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 100m*100m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

location of a taxi pick-up event needs to be protected when releasing information from the dataset.

We examine three larger cell sizes, 50 meters by 50 meters, 100 meters by 100 meters, and 200 meters by 200 meters. The noisy Bernoulli probabilities are combined into Poisson-Binomial distributions over the larger cells. We also compute the noisy counts $m_L^d(k)$ (Equation 2) and the noisy count probabilities $p_L^d(k)$ (Equation 3) for the larger cells.

We use the September 2013 New York Taxi data as validation. Again we take the 2pm–4:30pm period from Monday to Thursday in September, and break the taxi trip records into 5 minute intervals. Hence we obtain 630 repeated measurements in September for validation purpose. Using the noisy Bernoulli probabilities, Poisson-Binomial distributions, and the noisy count probabilities, we generate 100 copies of the synthetic counts in the larger cells. The maximum number of pickup events b in a cell ranges from 10 to 20 for 200 meters by 200 meters cells. b ranges from 4 to 6 for 100 meters by 100 meters cells. And b is around 1 to 3 for 50 meters by 50 meters cells. In one larger cell, the synthetic counts by each method

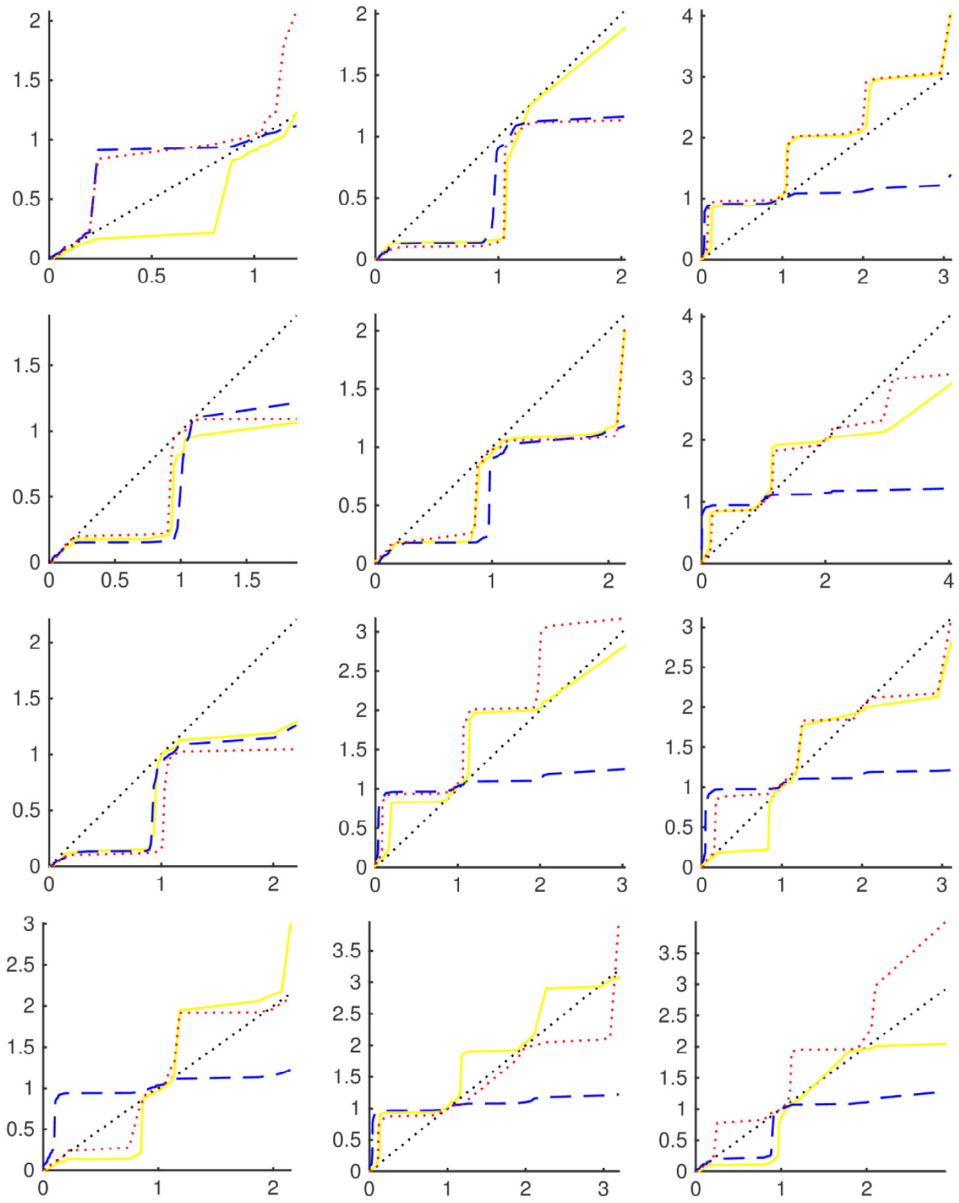

Fig. 3 – Comparing synthetic counts from three methods with the real counts in September. ϵ_{min} = 4. Spatial cell size 50 meters by 50 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 50m*50m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are ordered, and plotted against the 100 quantiles, $(i - 0.5)/100$ with $i = 1, \ldots, 100$, of the 630 repeated measurements.

Figs 1–3 have $\epsilon_{min} = 4$ for the smallest 5 meter by 5 meter cells; $\epsilon_{50} = 0.4$ for 50 meter by 50 meter cells; $\epsilon_{100} = 0.2$ for 100 meter by 100 meter cells; $\epsilon_{200} = 0.1$ for 200 meter by 200 meter cells. Figs 4–6 have $\epsilon_{min} = 2$ for the smallest 5 meter by 5 meter cells; $\epsilon_{50} = 0.2$ for 50 meter by 50 meter cells; $\epsilon_{100} = 0.1$ for 100 meter by 100 meter cells; $\epsilon_{200} = 0.05$ for 200 meter by 200 meter cells. On Figs 1–6, the Y axis stands for the ordered synthetic counts generated by the three methods. The X axis stands for the 100 quantiles of the 630 repeated measurements from September. We also have a 45 degree reference line going through the origin, created by plotting the 100 quantiles against themselves.

Synthetic counts generated from the noisy Bernoulli probabilities and the noisy Poisson-Binomial distributions are both more accurate than the noisy counts over different spatial cell sizes and different ϵ values. An interesting observation is that

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018

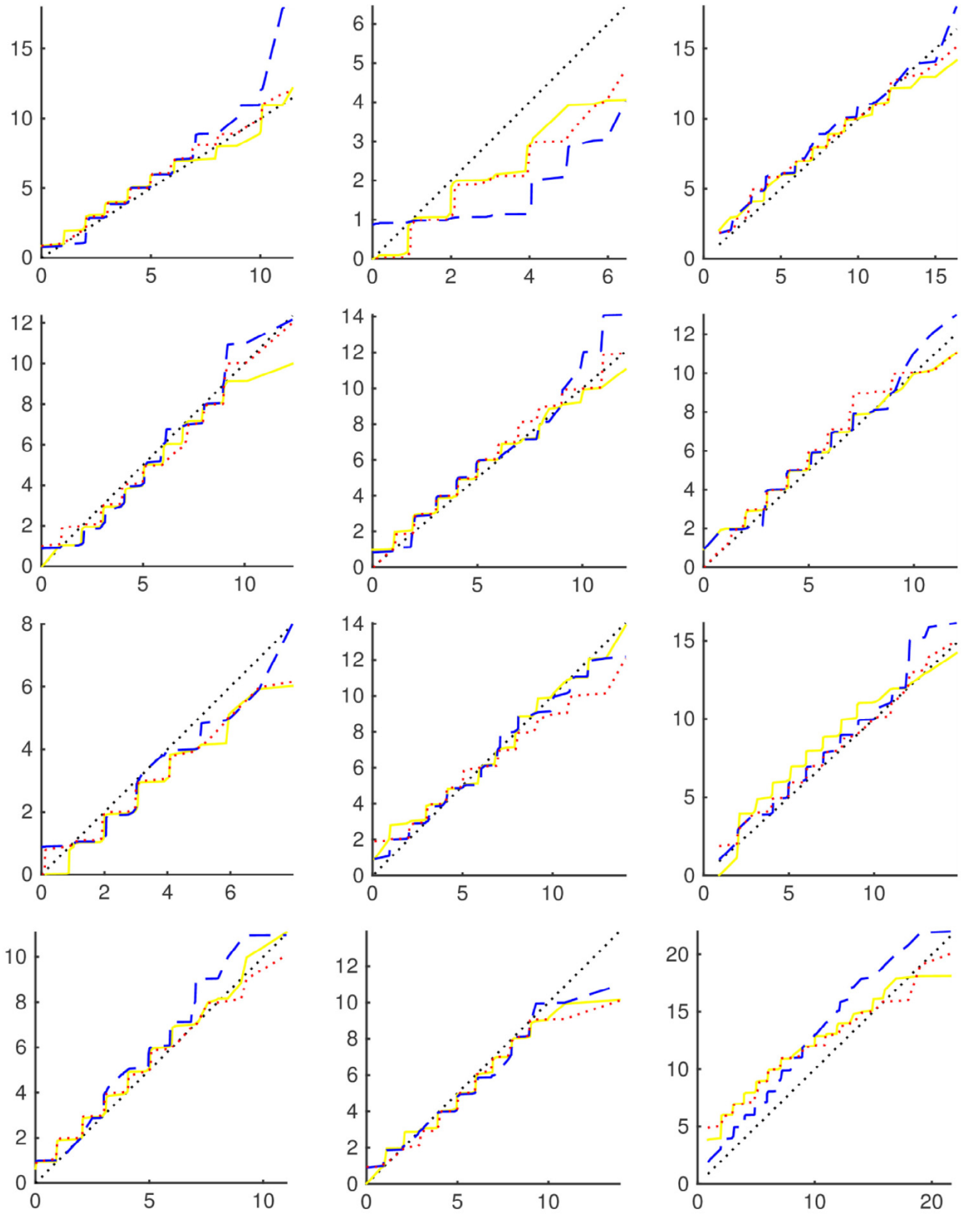

Fig. 4 – Comparing synthetic counts from three methods with the real counts in September. $\epsilon_{min} = 2$. Spatial cell size 200 meters by 200 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 200m*200m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the synthetic counts generated using the noisy count probabilities are less accurate in smaller cells, but quite close to the real data in the largest 200 meter by 200 meter cells. Note the variance of added noise is smaller in smaller cells. Hence the noisy count probabilities are less accurate under small injected noises. The 2nd plot in the top row of Fig 1 helps to explain this phenomenon. Cell size is not the most important factor for the noisy count probabilities. When the range of the count query is small, even if it is a large cell, the noisy count probability approach becomes less accurate. The noisy count probability approach has better accuracy in busy cells, with wider range of the count query.

5. Conclusion

In this paper we develop two methods to publish differential private spatial count probabilities, the Bernoulli method, and

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018

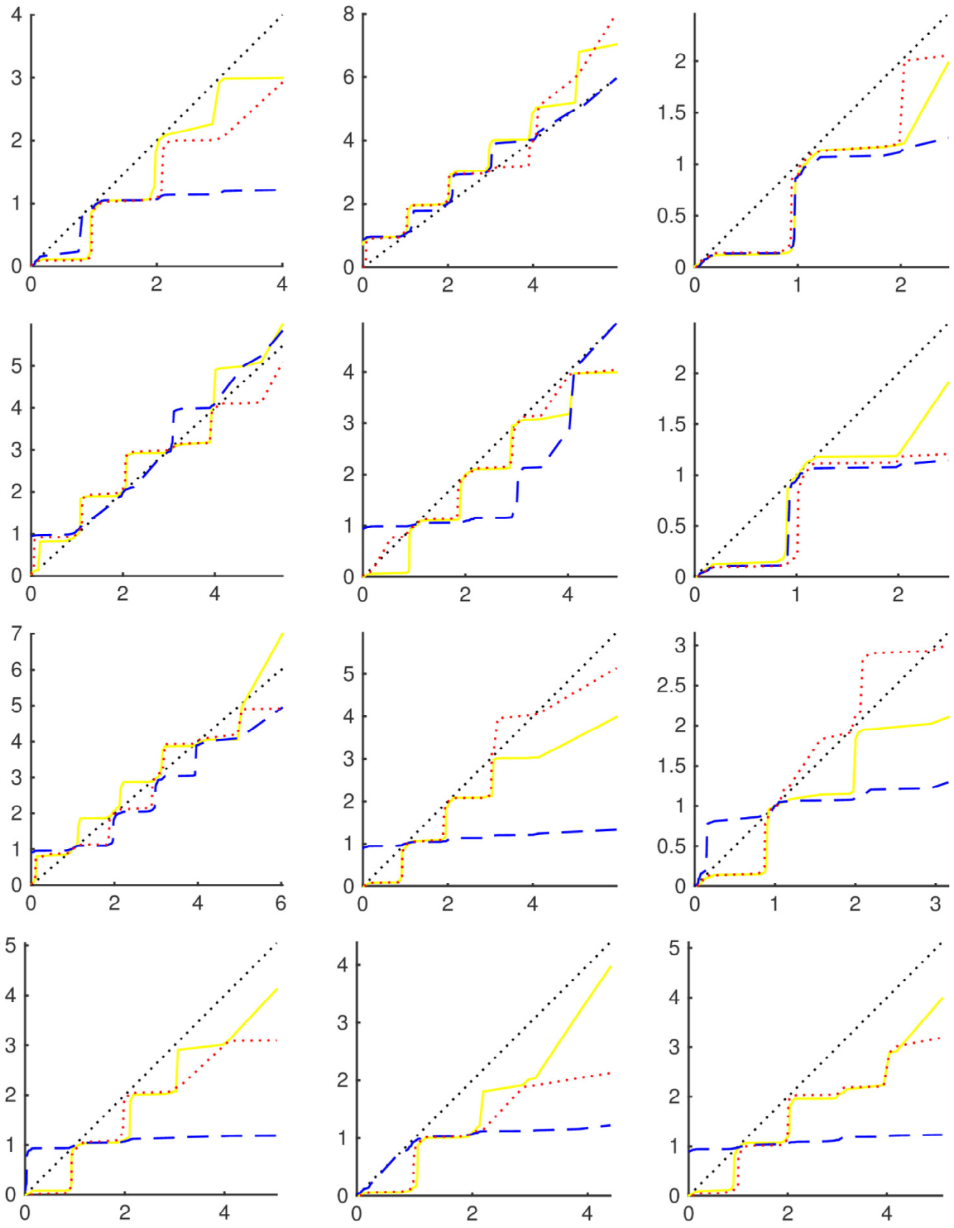

Fig. 5 – Comparing synthetic counts from three methods with the real counts in September. $\epsilon_{min} = 2$. Spatial cell size 100 meters by 100 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 100m*100m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the Poisson-Binomial method. Publishing the noisy Bernoulli probabilities computed from the smallest spatial cells dividing up a map region is the most flexible scheme, allowing users to freely move around the map, and to zoom in and zoom out at arbitrary locations. The noisy Poisson-Binomial distributions are computed from the noisy Bernoulli probabilities for larger spatial cells of high interest. Both methods are more accurate than the noisy count probabilities. In the experiments using New York Taxi data, synthetic counts generated using both methods match the real data accurately.

Acknowledgment

This work was supported in part by the Army Research Office [grant W911NF-17-1-0356], the National Institutes of Health [grant 1R01HG006844] and the National Science Foundation

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018

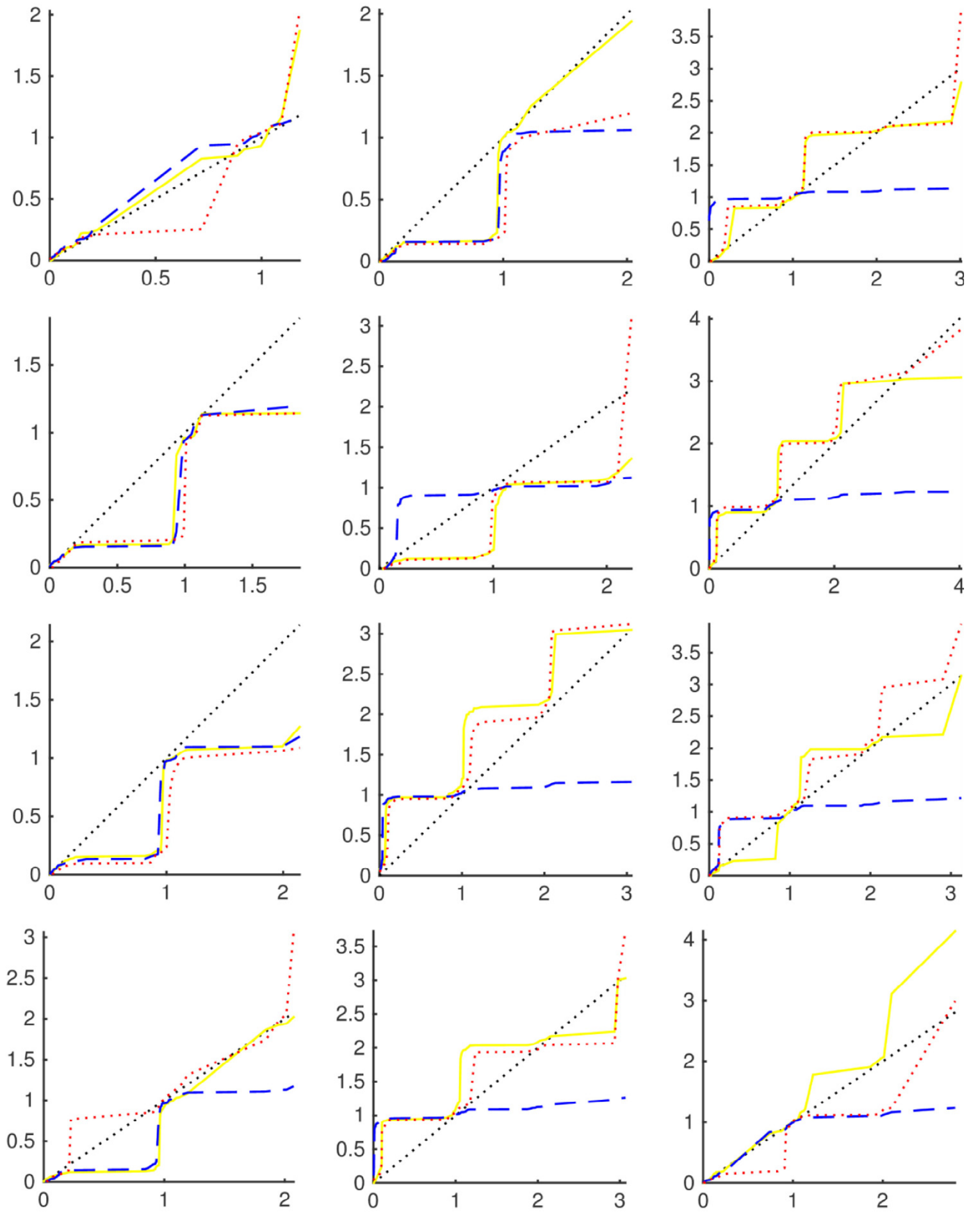

Fig. 6 – Comparing synthetic counts from three methods with the real counts in September. $\epsilon_{min} = 2$. **Spatial cell size 50 meters by 50 meters. Black dotted line is the 45 degree reference line going through the origin. Yellow solid line shows the counts generated using noisy Bernoulli probabilities. Red dotted line shows the counts generated by noisy Poison-Binomial distribution. Blue dashed line shows those generated by noisy count probabilities directly computed from the 50m*50m cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)**

[grants CNS-1111529, CNS-1228198, CICI-1547324, and IIS-1633331].

REFERENCES

Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy. JACM 2013;60(2):Article 12.

Donovan B, Work DB. Using coarse GPS data to quantify city-scale transportation system resilience to extreme events. arXiv preprint arXiv:1507.06011; 2015.

Dwork C. Differential privacy: a survey of results. In: International conference on theory and applications of models of computation, 1–19. Berlin: Heidelberg; 2008.

Dwork C, Smith A. Differential privacy for statistics: what we know and what we want to learn. J Priv Confid 2010;1(2):135–54.

Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: TCC, vol. 3876. 2006. p. 265–84.

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: privacy via distributed noise generation. In: Eurocrypt, vol. 4004. 2006. p. 486–503.

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018

Fernndez M, Williams S. Closed-form expression for the Poisson-Binomial probability density function. IEEE Trans Aerosp Electron Syst 2010;46(2):803–17.

Great-circle distance. Available from: https://en.wikipedia.org/wiki/Great-circle_distance.

Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Proceedings VLDB Endowment 2010;3(1–2):1021–32.

Hodges JL, Le Cam L. The Poisson approximation to the Poisson Binomial distribution. Ann Math Stat 1960;31(3):737–40.

Hong Y. On computing the distribution function for the Poisson Binomial distribution. Comput Stat Data Anal 2013;59:41–51.

Hsu J, Gaboardi M, Haeberlen A, Khanna S, Narayan A, Pierce BC, et al. Differential privacy: an economic method for choosing epsilon. In: IEEE 27th computer security foundations symposium. 2014. p. 398–410.

Kifer D, Lin BR. Towards an axiomatization of statistical privacy and utility. In: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. ACM; 2010. p. 147–58.

Latitude. Available from: https://en.wikipedia.org/wiki/Latitude.

Le Cam L. An approximation theorem for the Poisson binomial distribution. Pac J Math 1960;10(4):1181–97.

Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: theory meets practice on the map. In: IEEE 24th international conference on data engineering (ICDE 2008). 2008. p. 277–86.

Machanavajjhala A, Korolova A, Sarma AD. Personalized social recommendations: accurate or private. Proceedings VLDB Endowment 2011;4(7):440–50.

Mir DJ, Isaacman S, Cceres R, Martonosi M, Wright RN. Dp-where: differentially private modeling of human mobility. In: 2013 IEEE international conference on big data. 2013. p. 580–8.

New York City Taxi Trip Data (2010–2013). Available from: https://databank.illinois.edu/datasets/IDB-9610843.

Sarwate A, Monteleoni C, Chaudhuri K. Differentially private support vector machines, ArXiv: 0912.0071; 2009.

Wang Q, Zhang Y, Lu X, Wang Z, Qin Z, Ren K. RescueDP: real-time spatio-temporal crowd-sourced data publishing with differential privacy. In: The 35th annual IEEE international conference on computer communications (INFOCOM 2016). 2016. p. 1–9.

Wasserman L, Zhou S. A statistical framework for differential privacy. J Am Stat Assoc 2010;105(489):375–89.

Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms. IEEE Trans Knowl Data Eng 2011;23(8):1200–14.

Xu J, Zhang Z, Xiao X, Yang Y, Yu G, Winslett M. Differentially private histogram publication. VLDB J 2013;22(6):797–822.

Jun Jiang is a graduate student in Computer Science at Washington State University Vancouver. He is currently a research assistant at Distributed Systems Research LAB at WSUV, where he contributes himself in developing real-time machine learning tools for detecting malicious data injections in PMU data streams. He previously received a Masters degree in Industrial Engineering at Purdue University in 2016. His research interests lie in big data analytics, machine learning and smart grid.

Bowei Xi received her Ph.D in statistics from the Department of Statistics at the University of Michigan, Ann Arbor in 2004. She is an associate professor in the Department of Statistics at Purdue University. She was a visiting faculty in the Department of Statistics at Stanford University in summer 2007, and a visiting faculty at Statistical and Applied Mathematical Sciences Institute (SAMSI) from September 2012 to May 2013. Her research focuses on multidisciplinary work involving big datasets with complex structure from very different application areas including cyber security, Internet traffic, metabolomics, machine learning, and data mining. She has a US patent on an automatic system configuration tool and has filed another patent application for identification of blood based metabolite biomarkers of pancreatic cancer.

Murat Kantarcioglu is a Professor of Computer Science and Director of the UTD Data Security and Privacy Lab at The University of Texas at Dallas. He holds MS and PhD degrees in Computer Science from Purdue University. He is recipient of an NSF CAREER award and a Purdue CERIAS Diamond Award for academic excellence. He has been a visiting scholar at Harvard's Data Privacy Lab. Dr. Kantarcioglu's research focuses on creating technologies that can efficiently extract useful information from any data without sacrificing privacy or security. In addition, he focuses on using adversarial data mining techniques for fraud detection, cyber security and homeland security. His research has been supported by awards from NSF, AFOSR, ONR, NSA, and NIH. He has published over 150 peer-reviewed papers. His work has been covered by media outlets such as Boston Globe and ABC News, among others and has received three best paper awards. He is a senior member of both ACM and IEEE.

Please cite this article in press as: Jun Jiang, Bowei Xi, Murat Kantarcioglu, Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security (2017), doi: 10.1016/j.cose.2017.11.018