Privacy-Aware Collection of Aggregate Spatial Data

Abstract

Privacy concerns can be a major barrier to collecting aggregate data from the public. Recent research proposes negative surveys that collect negative data, which is complementary to the true data. This opens a new direction for privacy-aware data collection. However, the existing approach cannot avoid certain errors when applied to many spatial data collection tasks. The errors can make the data unusable in many real scenarios. We propose *Gaussian negative surveys.* We modulate data collection based on Gaussian distribution. The collected data can be used to compute accurate spatial distribution of participants and can be used to accurately answer range aggregate queries. Our approach avoids the errors that can occur with the existing approach. Our experiments show that we achieve an excellent balance between privacy and accuracy.

Keywords: Spatio-temporal Databases, Privacy, Aggregate Query, Negative Surveys, Geographic Information System

1. Introduction

Collecting aggregate data from users is important to many applications. A typical data collection shows a number of *categories* to participants and requires each participant to select one category. Such a data collection system normally maintains aggregate information, e.g., total number of participants, for each category. It is well known that privacy concerns of participants have a strong impact on data collection, especially when the participants need to answer sensitive questions [1, 2]. The effects of privacy concerns can be magnified when spatial data is collected due to the fact that spatial information usually relates to the physical presence of people. Failure to protect spatial privacy can result in serious harm, including physical assault, to individuals [3]. This can explain the fact that spatial privacy draws increasingly more attention from the public [4]. As the perceived privacy is a deciding factor for the participation in data collection [5], many participants may intentionally provide false data, e.g., report categories that does not apply to them, and may even refuse to provide any data [6]. Consequently, the collected data may contain significant errors in certain categories. In addition, the data may become more unusable when answering *range aggregate queries* as errors can lead to more unexpected results or hide problems of the data due to errors cancelling each other.

Although the quality of the collected data can be improved using background knowledge [7], we are interested in collecting high quality data in the first place. To achieve this, one needs to minimize the effects of the privacy concerns during data collection. Recent research proposes *negative surveys* [8], which are a type of privacy-aware data collection. Negative surveys collect *negative data* [9–11], which is complementary to the true data.

Researchers guarantee a strong privacy protection by *encouraging* participants to report categories that do *not* apply to them. These categories are called *negative categories.* In contrast to negative surveys, we call the traditional data collections that collect true data *positive surveys.*

We envisage a simplified scenario that is used throughout the paper: a transportation authority wishes to know the number of passengers who arrive at a central station from other stations on a railway route. In this example, there are seven stations where a passenger can get on a train (Figure 1). Aggregate data is maintained for each station.

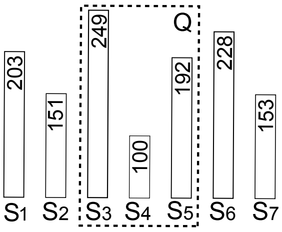

Figure 1: Responses to a public survey. S_1, …, S_7 are seven stations on a route to a central station. The numbers in the bars are the counts of participants from the corresponding stations. Q is the range of the query.

The query range Q shown in the example covers three adjacent stations, S_3, S_4, and S_5. The true answer to the query is 541. We compare two types of data collections in Figure 2. Assume that a participant starts a trip from station S_4. The participant has to report S_4 in a positive survey. However, in a negative survey the participant reports any station but S_4. One only knows that the reported station, for example S_2, is not the actual station where the participant started the trip in a negative survey. In this example, there are 6 possible stations, which safeguards the participant's location privacy.

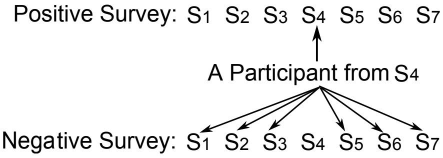
Positive Survey: S1 S2 S3 S4 S5 S6 S7
A Participant from S4
Negative Survey: S1 S2 S3 S4 S5 S6 S7

Figure 2: Comparison between a positive survey and a negative survey. S_1, …, S_7 are the same stations as in Figure 1.

We call the existing approach of collecting negative data *Uniform Negative Surveys (UNSs)* [8] because each category, except the true category, has an equal probability to be selected by participants. As the reported counts can be vastly different to the true counts, UNSs apply a technique to reconstruct the true data from the negative data. We detail the data reconstruction technique in Section 3.1. UNSs provide an accurate estimation of the true data if the ratio of participants to categories is high. Using the previous scenario of surveying passengers, we simulate a data collection with 10000 participants. We can observe that the estimated data from UNS are close to the true data (Figure 3).

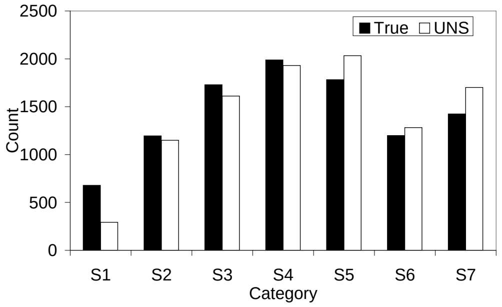

Figure 3: True data vs. estimated data from an UNS with 10000 participants. S_1, . . . , S_7 are the same stations as in Figure 1.

UNSs are not immune to errors during data reconstruction. We found that this problem is particularly severe when the ratio of participants to categories is low. For example, Figure 4 compares the true counts and the reconstructed counts from an UNS using the previous scenario of collecting aggregate data from passengers. The results are from a simulation with 100 participants. As the chart shows, the reconstructed data significantly deviate from the true data. When the data from adjacent categories are summed up for answering range aggregate queries, errors in individual categories may offset each other, if some of them cause over-counting while others cause under-counting. This may be desirable for an individual query but may lead to more significant problems in the future as it relies on erroneous data. For example, other queries may not show the same behavior leading to user confusion. Also, in many situations, the offset does not happen or the offset is insignificant compared to the answer. For example, using the data in Figure 4, if the query range covers S_2, S_3 and S_4, errors in those categories do not offset each other as all of them are lower than the true counts in corresponding categories. We can observe from Figure 4 that there are three categories with negative counts after reconstruction. The negative counts are unavoidable as shown in the original work on negative surveys. Those counts, which are wrong, further decrease the use of negative surveys. Thus, we propose GNSs, a method that does not generate negative counts.

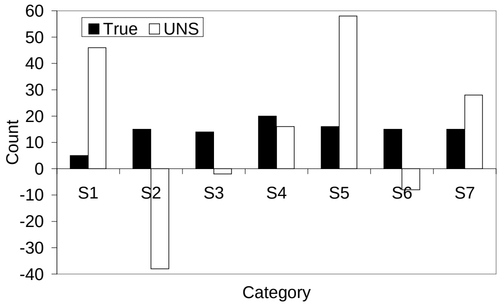

Figure 4: True data vs. estimated data from an UNS with 100 participants. S_1, . . . , S_7 are the same stations as in Figure 1.

In this paper, we propose *Gaussian Negative Surveys (GNSs)*. Different to UNSs, the reported categories are *not* uniformly chosen in GNSs. Instead, negative categories that are close to the true category are more likely to be selected than the negative categories far from the true category. Our work is focused on surveying geo-spatial data, which often shows strong correlation between adjacent locations in two-dimensional environments. One should not confuse the Gaussian distribution used by GNSs with the underlying distribution of the phenomena being surveyed, which does not have to follow Gaussian distribution. Our experimental results show the high accuracy levels achieved by GNSs for real geo-spatial data. We should also note that GNSs are applied to discrete categories because public surveys normally collect the data based on spatial decomposition with a certain resolution.

Our technique is particularly useful when aggregate spatial data is collected. Data collections involving spatial data may have hundreds or thousands of categories, each of which represents a certain location or area. The success of our approach lies in the fact that GNSs retain the correlation between adjacent categories in collected data because a majority of participants are expected to report the categories that are close to their true categories. As a range aggregate query normally covers a number of categories, a large portion of the reported counts within the query range would come from the participants who are actually in the range. Thus, the reported data can be used to approximate the true data in the query range. We should note that the participants can still achieve a high level of privacy protection when the data collection is modulated as above. This is because a category can be reported by participants from a number of nearby categories with similar probabilities. Consequently, similar to UNSs, it is still difficult for an adversary to derive the true category of a participant in GNSs. Our experimental results (Section 6) show that GNSs can achieve a high level of privacy while providing accurate answers to queries.

The contributions of this paper are:

- *We develop GNSs for collecting aggregate data while protecting individuals' privacy. GNSs are particularly suitable for collecting spatial data. GNSs are also significantly better than UNSs in answering range aggregate queries.*
- *We define a privacy metric for individuals in data collection.*
- *We analyze the factors that affect the privacy level and the accuracy level of GNSs.*
- *We compare GNSs with UNSs and a common data perturbation technique, Uniform Retention Replacement Perturbations [12], through comprehensive experiments. Our results show that GNSs achieve high privacy levels similar to the existing approaches but are significantly more accurate in solving queries.*

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 details UNSs and gives our analysis of the errors in UNSs. We present GNSs in Section 4. Privacy and accuracy metrics are introduced in Section 5. The experimental results are shown in Section 6. We conclude our paper in Section 7.

2. Related Work

2.1. Privacy Protection in Statistical Databases

Our work is related to privacy protection in statistical databases. Adam et al. [13] categorize the privacy-preserving approaches for statistical databases: *conceptual frame-*

work, *query restriction*, *data perturbation* and *output perturbation*. Conceptual frameworks address the security issues by defining the rules for constructing a database, such as avoiding the insertion of individual entities and building multi-dimensional tables with aggregated information. Query restriction protects the privacy by limiting the queries issued to a database. For example, successive queries may be blocked if new queries are highly overlapping with the existing query. Data perturbation protects the original data by modifying the original data or masking the original data [14, 15]. Different to data perturbation, output perturbation modifies the answer to a query when the query is solved based on original data. For example, the inclusion and exclusion of an entity in the answer to a query can be determined by a given probability. Negative surveys belong to *data perturbation* as the reported data is not the true data. Data perturbation can be further divided into *randomization*, *swapping*, *generalization* and *encryption*.

Randomization approaches distort the original data by adding a random value to the true value [12, 16, 17]. The randomizing parameter used for the distortion follows a certain distribution. Data swapping breaks the linkage between attributes by reordering a data matrix [18, 19]. Although the data is changed, it is still suitable for answering certain aggregate queries [20]. Data generalization builds a hierarchy of the summarized information [21, 22]. Information with the finest granularity is mapped to the lowest level in the hierarchy. Values in a higher level are generalized from a lower level. Data generalization techniques are suitable for achieving a certain level of anonymity. There are also techniques for solving queries using encrypted data without decryption. For example, Ge et al. proposed an approach to perform *SUM* and *AVG* queries on cipher text [23]. Another approach proposes a method to preserve the order of original values in encrypted data [24].

It is important to note that these approaches cannot avoid the collection of true data. For example, the randomization approaches allow that a certain percentage of the data is true. Our approach takes a different route from them as GNSs prohibit the collection of true data in the first place. This leads to a significantly stronger privacy protection for the participants. Compared to many of the approaches, GNSs are also substantially easier to implement as data reconstruction and encryption are not needed.

2.2. Spatial Data Access Methods

Research on aggregate spatial data is also highly related to our work. The work in this area follows two directions. The first direction focuses on hierarchical data structures, such as *aR-tree* [25], *CRB-tree* [26] and *aP-tree* [27]. Objects are sorted by their exact values, e.g., coordinates, in the lowest level of a hierarchy. Aggregate information, e.g., count of objects, is maintained at different levels of the hierarchy. The advantage of these approaches lies in the fact that data in child nodes does not need to be accessed when the spatial ranges of their parent nodes are contained in the query range. The second direction focuses on histogram-based techniques [28–32]. For example, *Euler histograms* provide an efficient way to answering range aggregate queries on point objects and rectangular objects [33, 34]. Many of the histogram-based approaches provide certain level of privacy protection as identities of participants are not required. GNSs provide an additional layer of privacy protection as participants do not report true categories in the first place.

2.3. Protection of Location Privacy

There is a body of work that concentrates on privacy protection in collecting and using spatial data. Shilton [35] investigates the privacy issues where the collection of location data involves the participation of individuals. The research highlights that there will be a higher level of privacy when participants can take control of their data, e.g., decide what can be collected and who can see the data. One recent paper stresses the importance of the balance between location privacy and the accuracy of the aggregated data [36]. When the precision of the collected data is high, e.g., the exact home addresses of patients are collected, the accuracy of the aggregated data is also high but location privacy is at its lowest level. Papadopoulos et al. [37] propose an approach to protect the location privacy of users who send *k Nearest Neighbor (kNN)* queries. The goal is to ensure that the original queries are processed in such a way that an adversary cannot distinguish the query location from other locations. Onsrud et al. [38] and Yeung and Hall [39] suggest that appropriate policies and education are vital to protecting individuals' privacy in spatial databases. Xu et al. [40] reveal that privacy intervention approaches, including compensation, industry self-regulation, and government regulation, have an impact on the perceived privacy in location-based services. *Spatial cloaking* adjusts the resolution of location information such that an individual cannot be distinguished from a number of other individuals [41, 42]. Obfuscation-based techniques degrade the quality of spatial information by maintaining a certain level of inaccuracy, imprecision or vagueness of spatial data [43, 44]. A technique similar to randomization methods for statistical databases has been utilized in studying spatial distribution of diseases [45]. *Geographic masking* protects location privacy by adding noise to geographic information [46]. Another recent research shows the use of transformed trajectories of moving objects that protects the privacy of individuals [47]. Mukherjee et al. [48] use perturbation and transformation to protect privacy in collecting spatial information. However, there is a lack of research that investigates the effects of location privacy protection techniques on answering range aggregate queries, which are a common way to use spatial data. In our work, we not only measure the privacy levels achieved by participants, but also the performance of our approach in answering range aggregate queries.

3. Preliminaries

3.1. Data Collection and Reconstruction in Uniform Negative Surveys

In UNSs, participants are shown mutually exclusive categories. For a participant, there is exactly one category, which is called the *positive category*; other categories are called the *negative categories* [8]. A participant randomly reports a negative category during data collection. The counts of participants for each category are maintained without knowing the true categories for the participants. An important assumption of UNSs is that the reported category is uniformly selected from the negative categories. Taking our previous example of collecting aggregate data from passengers (Figure 1), there are 7 categories in the questionnaire. That means 6 out of the 7 categories are negative categories for any participant. Each negative category has an equal chance, i.e., $\frac{1}{6}$, to be selected by a participant.

For a category j, the estimated count e_j can be calculated by finding the difference between the total number n of participants and an estimated number e'_j of participants, who *do not* belong to the category j, i.e., $e_j = n - e'_j$. e'_j can be obtained as follows. In a data collection of c categories, the chance that a participant from category i reports category j is $\frac{1}{c-1}$. On average, out of $c-1$ participants, one count goes to category j. Assume that there are r_j participants who report category j, one can calculate e'_j as $e'_j = (c-1) \cdot r_j$. Hence, e_j is given as

$$e_j = n - (c-1) \cdot r_j. \tag{1}$$

3.2. Magnification of Errors

The original work on negative surveys assumes that UNSs can work well as long as participants report negative categories based on a perfectly uniform distribution. However, the original work on UNSs overlooked a particular type of errors during data reconstruction. As shown in the previous work on negative surveys [8] and our example (Figure 4), UNSs may give highly inaccurate statistics. We argue that the magnification of certain errors during data reconstruction can significantly affect the accuracy of UNSs. Let us denote the number of categories as c and the true count for category i as t_i. Theoretically, the number of participants who belong to category i but report category j is $\frac{t_i}{c-1}$, which is normally a real number. As counting is based on integers, the actual number is rounded from the original value. Hence, the actual number is either $\lfloor \frac{t_i}{c-1} \rfloor$ or $\lceil \frac{t_i}{c-1} \rceil$. The gap between the rounded value and the original value, which is $\frac{t_i}{c-1} - \lfloor \frac{t_i}{c-1} \rfloor$ or $\lceil \frac{t_i}{c-1} \rceil - \frac{t_i}{c-1}$, causes errors. We consider such errors as unavoidable.

Due to the reconstruction technique (Formula 1), the errors are magnified by a factor of $c-1$. For category j, the magnified errors caused by the participants from category i is either

$$Er_{i,j} = (c-1) \cdot (\frac{t_i}{c-1} - \lfloor \frac{t_i}{c-1} \rfloor)$$

or

$$Er_{i,j} = (c-1) \cdot (\frac{t_i}{c-1} - \lceil \frac{t_i}{c-1} \rceil).$$

As $\frac{t_i}{c-1} - \lfloor \frac{t_i}{c-1} \rfloor$ is between $[0, 1)$ and $\frac{t_i}{c-1} - \lceil \frac{t_i}{c-1} \rceil$ is between $(-1, 0]$, the upper/lower bound of $Er_{i,j}$ is $\pm(c-1)$. The total magnified errors at category j, Er_j, is:

$$Er_j = \sum_{i \in S_{\overline{j}}} Er_{i,j},$$

where $S_{\overline{j}} = \{1, \ldots, c\} \setminus \{j\}$. Although $Er_{i,j}$ from a different category i might offset each other to a certain extent when some cause over-counting while others cause under-counting, in practice, the chance that errors are canceled each other out is very slim or it is not known whether such cancellations can be really usable. In the worst case, all $Er_{i,j}$ cause over-counting or under-counting. As there are $c-1$ categories in set $S_{\overline{j}}$, the upper/lower bound of Er_j is $\pm(c-1)^2$. We should note that the errors can be relatively insignificant when there are a high number of participants per category, e.g., Figure 3 compared with Figure 4. This is because when a category contains a large number of participants, the errors may be equal to a small percentage of the true count. However, it is usually unrealistic to have a particularly high number of participants in a data collection. In general, data collections, e.g., public surveys, are done to achieve the opposite, i.e., estimating the behavior of a larger population from a small sample.

Taking our previous example of counting passengers at railway stations (Section 1), the number of participants is usually limited to a few hundred passengers during a regular period of data collection. It is also unrealistic to assume that there are always a low number of categories in a data collection. For example, a survey related to geographic information may have hundreds or thousands of categories, e.g., localities or stations in a city.

4. Gaussian Negative Surveys

We call our approach *Gaussian negative surveys* because the probabilities for selecting negative categories follow a Gaussian distribution centered at the true category. Due to this characteristic, the majority of true counts for a category is spread into the categories that are close to the category. Using the same example of collecting aggregate data from passengers (Figure 1), we show the probabilities that participants in positive category i report negative category j in Figure 5, where i and j correspond to station S_i and station S_j respectively. Note that some bars cannot be seen as the probabilities matching to them are zero. Figure 5 shows that the probability of reporting a negative category decreases when the negative category is further away from the positive category. Taking the example shown in Figure 1, the percentage of participants who belong to S_4 and report station S_j is 9%, 16.7%, 24.3%, 0%, 24.3%, 16.7% and 9% when j varies from

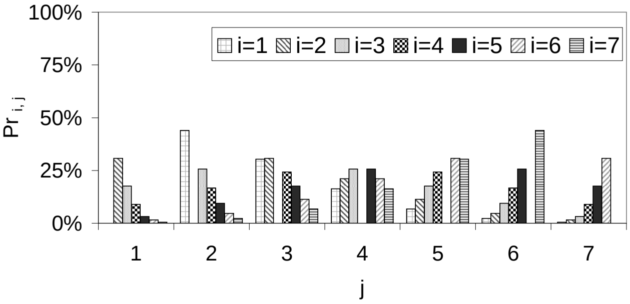

Figure 5: Probability that a participant in positive category i reports negative category j.

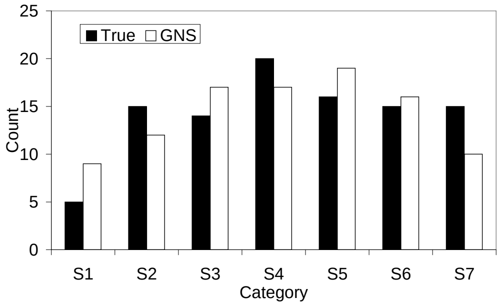

Figure 6: True counts vs. collected counts from a GNS with 100 participants. $S_1, \ldots, S_7$ are the same stations as in Figure 1.

Table 1: Probabilities sampled from density function of Gaussian distribution $f(j; i, \sigma^2)$ with $\sigma = 2$.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$	$j = 7$
$i = 1$	19.9%	17.6%	12.1%	6.5%	2.7%	0.9%	0.2%
$i = 2$	17.6%	19.9%	17.6%	12.1%	6.5%	2.7%	0.9%
$i = 3$	12.1%	17.6%	19.9%	17.6%	12.1%	6.5%	2.7%
$i = 4$	6.5%	12.1%	17.6%	19.9%	17.6%	12.1%	6.5%
$i = 5$	2.7%	6.5%	12.1%	17.6%	19.9%	17.6%	12.1%
$i = 6$	0.9%	2.7%	6.5%	12.1%	17.6%	19.9%	17.6%
$i = 7$	0.2%	0.9%	2.7%	6.5%	12.1%	17.6%	19.9%

1 to 7. Similarly, a category is more likely to be reported by participants from nearby categories than far away categories. Taking the same example in Figure 1, the percentage of participants who belong to station S_i and report station S_4 are 16.3%, 21.1%, 25.7%, 0%, 25.7%, 21.1% and 16.3% when i varies from 1 to 7. Let us assume that the true counts of participants in seven categories are 5, 15, 14, 20, 16, 15 and 15. Figure 6 compares the true data and the collected data using GNS for 100 participants. Compared with the performance of UNSs for the same setting (Figure 4), GNSs give a significantly better estimation of the true statistic.

We use the collected counts in GNSs to solve aggregate queries without data reconstruction as in UNSs. GNSs' answers to aggregate queries are usually close to the true answers due to two reasons. First, as true counts are clustered around positive categories, a major portion of the true counts within a query range are likely to be covered by the range. Second, although no counts are reported to positive categories, the missing counts for the positive categories are partially offset by the reported counts from other categories. We should note that GNSs may not give accurate answers if the true counts in adjacent categories are vastly different to each other. However, such situations rarely happen for spatial scenarios. An important observation of spatial information is the *first law of geography* [49]: everything is related to everything else, but near things are more related than distant things. One often expects a strong correlation of spatial data between adjacent categories. For example, the counts of patients in adjacent street blocks during an epidemic are usually close to each other. Hence, GNSs are unlikely to cause significant errors in approximating the spatial distribution. This makes GNSs particularly suitable for collecting aggregate spatial data.

We formulate the selection of negative categories as follows. We denote $f(j; i, \sigma^2)$ as the continuous probability density function for a Gaussian distribution, which is centered at a positive category i with a standard deviation σ. We use the function sampled at j as the probability that participants in category i report category j. Table 1 shows the sampled probabilities based on the previous example of collecting aggregate data about passengers from 7 stations.

Since negative surveys require that participants do not report positive categories, the probability of reporting positive category i is zero. Therefore, the probabilities for selecting negative categories need to be adjusted such that the sum of the adjusted probabilities is 1 (or 100% using Table 2's notation). Assuming $S_{\bar{i}} = \{1, \ldots, c\} \setminus \{i\}$, the

Table 2: Adjusted probabilities $Pr_{i,j}$.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$
$i=1$	0%	44%	30.2%	16.2%	6.8%	2.3%	0.5%
$i=2$	30.7%	0%	30.7%	21.1%	11.3%	4.7%	1.5%
$i=3$	17.8%	25.8%	0%	25.8%	17.8%	9.6%	3.2%
$i=4$	9%	16.7%	24.3%	0%	24.3%	16.7%	9%
$i=5$	3.2%	9.6%	17.8%	25.8%	0%	25.8%	17.8%
$i=6$	1.5%	4.7%	11.3%	21.1%	30.7%	0%	30.7%
$i=7$	0.5%	2.3%	6.8%	16.2%	30.2%	44%	0%

Table 3: Number of participants reporting a certain category. The last row shows r_j calculated from Formula 3 with j varies from 1 to 7. Each r_j is the sum of $t_i \cdot Pr_{i,j}$ with i varies from 1 to 7, which are also shown in the table.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$
$t_1 \cdot Pr_{1,j}$	0	2	1	1	0	0	0
$t_2 \cdot Pr_{2,j}$	5	0	5	3	2	1	0
$t_3 \cdot Pr_{3,j}$	2	4	0	4	2	1	0
$t_4 \cdot Pr_{4,j}$	2	3	5	0	5	3	2
$t_5 \cdot Pr_{5,j}$	0	2	3	4	0	4	3
$t_6 \cdot Pr_{6,j}$	0	1	2	3	5	0	5
$t_7 \cdot Pr_{7,j}$	0	0	1	2	5	7	0
r_j	9	12	17	17	19	16	10

adjusted probability that participants in positive category i report negative category j, $Pr_{i,j}$, is computed as

$$Pr_{i,j} = \begin{cases} \dfrac{f(j; i, \sigma^2)}{\sum_{k \in S_{\bar{i}}} f(k; i, \sigma^2)} & i \neq j, \\ 0 & i = j. \end{cases} \tag{2}$$

Based on the sampled probabilities shown in Table 1, we show the adjusted probabilities $Pr_{i,j}$ in Table 2. When a GNS is modulated based on the adjusted probabilities, the total number of participants who report category j, r_j, is

$$r_j = \sum_{i=1}^{c} t_i \cdot Pr_{i,j}, \tag{3}$$

where t_i is the true count for category i. Based on the previous example in Figure 1, we show r_j and the intermediate values for computing r_j in Table 3.

Let $g_{(a,b)}$ denote the answer to an aggregate query with the query range from category a to category b. The GNS's answer to the query is

Table 4: Answers to aggregate queries. In the table, (i, j) denotes the range of a query that asks for the total counts from category i to category j. We use $g_{(i,j)}$ to denote the answer of GNSs to the query. We use $t_{(i,j)}$ to denote the true answer to the query. Parameter s is the query size in terms of the number of categories in a query range. For example, when $s = 3$, an aggregate query covers three consecutive categories, i.e., the query range could be $(1, 3)$, $(2, 4)$, etc.

$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$
$t_{(1,1)} = 5$	$t_{(1,2)} = 20$	$t_{(1,3)} = 34$	$t_{(1,4)} = 54$	$t_{(1,5)} = 70$	$t_{(1,6)} = 85$	$t_{(1,7)} = 100$
$t_{(2,2)} = 15$	$t_{(2,3)} = 29$	$t_{(2,4)} = 49$	$t_{(2,5)} = 65$	$t_{(2,6)} = 80$	$t_{(2,7)} = 95$	
$t_{(3,3)} = 14$	$t_{(3,4)} = 34$	$t_{(3,5)} = 50$	$t_{(3,6)} = 65$	$t_{(3,7)} = 80$		
$t_{(4,4)} = 20$	$t_{(4,5)} = 36$	$t_{(4,6)} = 51$	$t_{(4,7)} = 66$			
$t_{(5,5)} = 16$	$t_{(5,6)} = 31$	$t_{(5,7)} = 46$				
$t_{(6,6)} = 15$	$t_{(6,7)} = 30$					
$t_{(7,7)} = 15$						
$g_{(1,1)} = 9$	$g_{(1,2)} = 21$	$g_{(1,3)} = 38$	$g_{(1,4)} = 55$	$g_{(1,5)} = 74$	$g_{(1,6)} = 90$	$g_{(1,7)} = 100$
$g_{(2,2)} = 12$	$g_{(2,3)} = 29$	$g_{(2,4)} = 46$	$g_{(2,5)} = 65$	$g_{(2,6)} = 81$	$g_{(2,7)} = 91$	
$g_{(3,3)} = 17$	$g_{(3,4)} = 34$	$g_{(3,5)} = 53$	$g_{(3,6)} = 69$	$g_{(3,7)} = 79$		
$g_{(4,4)} = 17$	$g_{(4,5)} = 36$	$g_{(4,6)} = 52$	$g_{(4,7)} = 62$			
$g_{(5,5)} = 19$	$g_{(5,6)} = 35$	$g_{(5,7)} = 45$				
$g_{(6,6)} = 16$	$g_{(6,7)} = 26$					
$g_{(7,7)} = 10$						

$$g_{(a,b)} = \sum_{j=a}^{b} r_j. \tag{4}$$

Based on the previous example, we compare the true answers and GNS's answers to aggregate queries in Table 4, which shows that GNSs give a good approximation of the true answers.

Accuracy levels achieved by GNSs are affected by two factors: *the shape of the distribution for selecting negative categories* and *the true counts near the boundary of the query range*. When the distribution is highly clustered around the true category, i.e., the standard deviation of the distribution is low, a majority of the participants in a query range are highly likely to report categories within the range. In this situation, miscounting mainly comes from the participants who are near the boundary of the query range. Under-counting happens when participants within the query range report categories outside the range. Over-counting happens when participants outside the query range report categories within the query range. The two types of miscounting may offset each other to a certain extent. We should note that the privacy level can be adversely affected when the distribution is highly clustered around the true category. Hence, one needs to adjust the distribution to achieve a good balance between accuracy and privacy. Our experimental results show that GNSs can achieve a high accuracy level and a high privacy level at the same time. The relationship between the two factors and the answers to range aggregate queries are presented in Theorem 1.

Theorem 1. *Let (a, b) be the range of a range aggregate query that asks for the cumulative count from category 'a' to category 'b'. Let $t_{(a,b)}$ be the true answer to the query and*

$g_{(a,b)}$ be the answer given by a GNS. Let t_i be the true count for category i. Let σ be the standard deviation for the GNS. We have $\lim_{\sigma\to 0}\dfrac{g_{(a,b)}}{t_{(a,b)}} = 1 + \dfrac{(t_{a-1}-t_a)+(t_{b+1}-t_b)}{2\cdot t_{(a,b)}}$.

Our theorem shows why GNSs can perform well for spatial data. As the spatial distribution of participants usually shows a consistent trend within a certain distance, the value of $(t_{a-1}-t_a)$ and $(t_{b+1}-t_b)$ are usually small. Different to these values, $t_{(a,b)}$ is less likely to be small because range aggregate queries often cover a large number of categories, which may contain a considerable number of participants. Hence, $\frac{(t_{a-1}-t_a)+(t_{b+1}-t_b)}{2\cdot t_{(a,b)}}$ is usually low and answer of GNSs is usually close to the true answer.

Proof of Theorem 1. For a positive category i, we assume that the probability density function of Gaussian distribution is $f(x;i,\sigma^2)$. We also assume that there are c categories. We define $A=\{1,\ldots,c\}\setminus\{i-1,i,i+1\}$.

$$\begin{aligned}\lim\nolimits_{\sigma\to 0} Pr_{i,i+1} &= \lim\nolimits_{\sigma\to 0}\frac{f(i+1;i,\sigma^2)}{\sum_{x\in A} f(x;i,\sigma^2)+f(i+1;i,\sigma^2)+f(i-1;i,\sigma^2)}\\ &= \frac{f(i+1;i,\sigma^2)}{f(i+1;i,\sigma^2)+f(i-1;i,\sigma^2)}.\\ &= 0.5\end{aligned}$$

Let $t_{i,j}$ be the number of participants who belong to category i but report category j. We get

$$\lim_{\sigma\to 0} t_{i,i+1} = \lim_{\sigma\to 0} t_i\cdot Pr_{i,i+1} = \frac{1}{2}t_i.$$

Similarly, we can prove that

$$\lim_{\sigma\to 0} t_{i,i-1} = \frac{1}{2}t_i.$$

Hence, the true count of a category is distributed evenly in two adjacent categories when σ is very small. That is,

$$\lim_{\sigma\to 0} g_i = \frac{1}{2}t_{i-1}+\frac{1}{2}t_{i+1}$$

We can deduct that

$$\begin{aligned}\lim\nolimits_{\sigma\to 0} g_{(a,b)} &= \textstyle\sum_{i=a}^{b}\frac{1}{2}(t_{i-1}+t_{i+1})\\ &= \tfrac{1}{2}\cdot(t_{a-1}+t_{a+1}+t_a+t_{a+2}+t_{a+1}+t_{a+3}+\ldots+t_{b-3}+t_{b-1}\\ &\quad +t_{b-2}+t_b+t_{b-1}+t_{b+1})\\ &= \tfrac{1}{2}\cdot(t_{a-1}+t_{b+1})-\tfrac{1}{2}(t_a+t_b)+t_{(a,b)}\end{aligned}$$

Hence,

$$\lim_{\sigma\to 0}\frac{g_{(a,b)}}{t_{(a,b)}} = 1+\frac{(t_{a-1}-t_a)+(t_{b+1}-t_b)}{2\cdot t_{(a,b)}}.$$

Table 5: A table randomly generated for a participant.

Your True Answer	1	2	3	4	5	6	7
You Could Report	2 3 4	5 1 3	1 4 2	3 5 6	7 4 6	7 5 4	6 5 4

Although GNSs can be applied to many types of data, the technique is particularly suitable for geo-spatial data because a common characteristic of geo-spatial data is the strong correlation between categories. Due to the correlation, counts in adjacent categories do not change significantly. Different to UNSs, GNSs can show the trend of changes across a potentially large number of categories in geo-spatial data. We conducted comprehensive experiments on synthetic and real geo-spatial data. The results highlight the advantages of GNSs in surveying this type of data.

4.1. Implementing GNSs

One essential characteristic of GNSs is that participants report different categories with different probabilities. We present a method to control the probabilities in real data collections.

The method can be implemented in an electronic device, e.g., a tablet computer. Upon joining a survey, a participant is shown a table on the screen of the device. The table is randomly generated in real-time. The table shows all positive categories, each of which is attached with a subset of negative categories. The participant can report any of the given negative categories corresponding to her positive category. To enhance the trust of the system, one can add a feedback component to the system. That is, if a participant is not satisfied with the given selection, she is allowed to re-initiate the table-generating process until she is satisfied.

For a positive category i, the subset of negative categories are generated based on $Pr_{i,j}$, which is the probability that participants from category i report category j. For example, if $Pr_{3,4}$ is 0.4, then category 4 has a 40% chance to appear in the subset for positive category 3. We should note that the system will ensure that all negative categories in a subset are distinct to each other. Let us see an example based on the scenario of surveying passengers described in the first section. As shown in Table 5, a participant who belongs to category 3 could report category 1, 2 or 4. For scenarios where there is no electronic device to generate the table in real-time, a surveyor can apply the idea to a paper-based guidance. That is, the surveyor generates and prints a sufficient number of tables before the survey.

4.2. Controlling Privacy

We present two approaches to controlling the privacy levels of GNSs. Our first approach controls privacy by customizing the implementation of GNSs. As shown in Section 4.1, a participant is given a subset of negative categories for each positive category. The size of the subset affects the level of perceived privacy. Participants would feel that privacy is at the lowest level if there is only one negative category for a positive category.

This is because reporting a negative category would let an adversary find out the positive category if the adversary knows the one-to-one mapping between the two types of categories. When there is a higher number of negative categories attached to a positive category, there would be a higher level of perceived privacy. Therefore, surveyors can control the privacy level by changing the size of the subset of negative categories attached to a positive category.

Different to the first approach, the second approach controls privacy by determining the systemic parameters of surveys. To quantify the effects of the systemic parameters, we adapt the concept of *k-anonymity* [50]. In our case, k-anonymity means that a participant, who reports category i, cannot be distinguished from other $k-1$ participants who also report category i. By adjusting certain settings in a GNS, one can control the degree of k-anonymity, and thus control the privacy level of GNS. Assume that there are c categories and n participants. We denote the k-anonymity for category j as ka_j. We also denote the standard deviation used in Gaussian distribution as σ. As we do not know the actual counts of participants in each category, we assume that the true counts are uniformly distributed in the categories. Hence, there are $\frac{n}{c}$ participants in each category. For any category i that is not equal to j, the number of participants who are in category i but report category j is $Pr_{i,j} \cdot \frac{n}{c}$. We have the following formula.

$$ka_j = \sum_{i \neq j} Pr_{i,j} \cdot \frac{n}{c}$$

Surveyors of GNSs can control the range of k-anonymity by applying different combinations of c, n and σ. Let us see an example based on the previous scenario of surveying passengers (see Section 1) where c is 7. Assume that the surveyor sets σ to 2. Table 2 shows the probabilities $Pr_{i,j}$ under these settings. When there are 100 participants, i.e., $n = 100$, ka_j will be 9, 15, 17, 18, 17, 15 and 9, when j varies from 1 to 7. Hence, the surveyor controls the k-anonymity between 9 and 18. In extreme case, when σ is large, the shape of the Gaussian distribution is close to the uniform distribution. In that case, the k-anonymity level will be either 14 or 15 in our example and thus approximately be constant.

5. Measuring Privacy and Accuracy

5.1. Privacy

We define a measure of privacy in GNSs. We do not measure the actual privacy level achieved by individuals based on k-anonymity. Anonymity-based metrics evaluate the privacy levels when people can report true sensitive information but want to be indistinguishable from others. Survey participants, however, usually want to hide sensitive information in the first place. We define a metric that estimates the privacy levels of an individual without knowing the true data. We measure privacy based on the probability that a positive category is derived from a reported category. If the true category has a high probability to be derived, the privacy of the participant is at high risk. Otherwise, the privacy is well protected. In our experiments, we calculate the average privacy level gained by all participants using this measure.

Let us consider how a positive category can be derived from a reported category. Assuming there are c categories, a participant who reports category j belongs to one of the categories in set $S_{\overline{j}} = \{1, \ldots, c\} \setminus \{j\}$. As the true statistic is unknown, a data collector has to assume that the true counts are uniformly distributed in c categories. Let $Pr_{i,j}$ denotes the probability that participants in category i report category j. The probability that a participant actually belongs to category i is

$$\frac{Pr_{i,j}}{\sum_{k \in S_{\overline{j}}} Pr_{k,j}}.$$

We define the privacy level based on this probability:

$$Privacy_{i,j} = 1 - \frac{Pr_{i,j}}{\sum_{k \in S_{\overline{j}}} Pr_{k,j}}. \tag{5}$$

Let us see an example based on the probabilities shown in Figure 5. Assume that a participant in category 3 reports category 1. Since $Pr_{k,1}$ is 30.7%, 17.6%, 9%, 3.2%, 1.6% and 0.5% when k varies from 2 to 7, the privacy level of the participant is

$$1 - \frac{0.176}{0.307 + 0.176 + 0.09 + 0.032 + 0.016 + 0.005} = 72\%.$$

5.2. Accuracy

We measure accuracy level from three aspects. First, we use the standard *Root Mean Square Error (RMSE)* to measure the absolute value of errors. The previous work, UNSs, uses a customized metric based on RMSE [8]. For the ith query in q queries, e_i denotes the estimated answer from negative surveys and t_i denotes the true answer to the query. RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{q}(e_i - t_i)^2}{q}}. \tag{6}$$

Second, we measure *Relative Accuracy (RA)* based on the ratio of error to the true answer. We denote the true answer as t and the answer given by negative surveys as n. Formula 7 presents our definition of RA. This measurement shows the relative significance of errors. For example, let us assume that a true answer is 10000 and the answer given by a negative survey is 9500. Although the negative survey has an error of 500, the error is not significant compared to the true count, i.e., $RA = 95\%$.

$$RA = \begin{cases} 1 - \dfrac{|n-t|}{t} & when \quad |n-t| <= t, \\ \\ 0 & when \quad |n-t| > t. \end{cases} \tag{7}$$

The above two metrics are for evaluating the accuracy of answers to range aggregate queries. In our experiments, we also measure the similarity between the true spatial distribution and the collected spatial distribution using *Kolmogorov-Smirnov test (KS-test)* [51, 52]. Different to RMSE and RA, we consider all categories for KS-tests, not only the categories in the query range. KS-test gives D-value, which is a normalized distance between two distributions. The range of D-value is between 0 and 1. A small D-value means two distributions are close to each other. For example, the D-value for the two distributions in Figure 4 is 0.4286 while the D-value for the two distributions in Figure 6 is 0.2857.

6. Experiments

We compare our proposed approach, GNSs, with the existing approach of negative surveys, UNSs. We also compare GNSs with a common data perturbation technique, Retention Replacement Perturbations [12]. This approach works as follows. During a survey, a random value is generated for each participant. If the value is below a pre-determined retention threshold, the true category of the participant is collected. Otherwise, one of the categories (including the true category) is randomly selected based on an uniform distribution. Similar to UNSs, the collected data needs to be put through a reconstruction process before solving queries. Due to the use of uniform distribution, we call this approach Uniform Retention Replacement Perturbations (URRPs). The original work on URRPs shows that a lower retention threshold leads to a higher privacy level. As privacy is of priority in our research, we set the retention threshold to a very low value, 0.01. As shown in the experimental results, this setting allows URRPs to achieve high privacy levels similar to UNSs.

We conduct experiments in one dimensional and two dimensional settings. The one dimensional settings simulate data collections that are similar to our example of collecting aggregate data about passengers on a route (Figure 1). The two dimensional settings simulate data collections in which the number of categories is high, such as surveys of patients in a country by region. Besides the synthetic spatial data, we also evaluate all approaches using a set of real 2D data. We use various settings through out the experiments. For each setting, we run the same tests for 100 times and average the results. In each run, we first generate a true data set. We then generate the collected data based on the true data set. For each of the surveys, we evaluate the similarity between the true distribution and the collected distribution using KS-tests. The average privacy levels are also calculated. In addition, we process 100 range aggregate queries using true data and the collected data from the two negative surveys. The size and position of the query ranges are randomly determined using an uniform distribution. Based on the answers to the queries, RMSE and the averaged RA are calculated. Table 6 presents the values of experimental parameters. There are two types of settings, one is *default settings*, another is *alternative settings*. Each experiment varies one of the four parameters based on the alternative settings while keeping the other three parameters to the default settings. The parameters are defined as follows.

- **Number of Categories**: This is the number of stations on a route (Figure 1). For two-dimensional settings, this parameter is the number of cells in a grid partitioning of the space.
- **Standard Deviation of GNSs**: This is the parameter σ used in Formula 2. It determines the shape of the Gaussian distribution in GNSs. A low value means that most participants would select negative categories that are close to the true categories, and vice versa.

Table 6: Experimental settings.

	Default Settings	Alternative Settings
Number of Categories	20 (1D), 20 × 20 (2D)	5 - 35 (1D), 5 × 5 - 35 × 35 (2D)
Standard Deviation of GNSs	2	0.5 - 3.5
Query Size	25%	5% - 45%
Number of Participants	1000	400 - 1600

- **Query Size**: This is measured as the ratio between the number of categories in query range to the total number of categories.
- **Number of Participants**: This is the total number of participants in a data collection. We should note that this parameter is fixed for real 2D datasets.

6.1. Synthetic One Dimensional Datasets

For one dimensional settings, the *true* counts are generated from one of two common distributions: Gaussian distribution and uniform distribution. We set the standard deviation of Gaussian distribution to one fifth of the number of categories. This is for simulating common scenarios where counts of participants do not change dramatically and remain constant across adjacent geographic areas. When the true distribution is Gaussian, we label the results from UNSs, GNSs and URRPs as UNS(G), GNS(G) and URRP(G), respectively. When the true distribution is uniform, we label the results from UNSs, GNSs and URRPs as UNS(U), GNS(U) and URRP(U), respectively.

6.1.1. Number of Categories

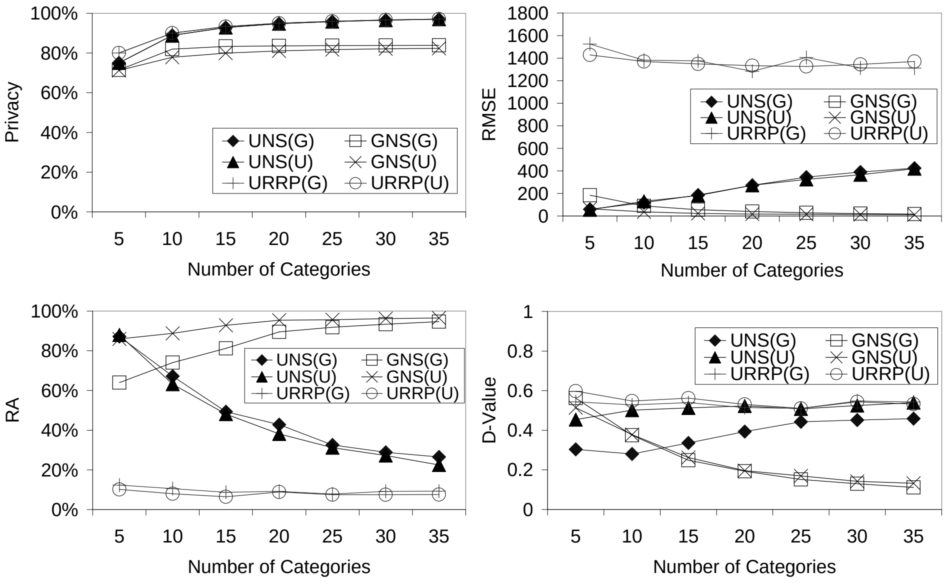

Figure 7: Effects of the number of categories in synthetic one dimensional settings.

We observe the advantages of GNS over UNS with the increase in the number of categories (Figure 7). All approches achieve 80% and higher privacy levels when there are 15 or more categories. GNS's RMSE decreases with the growth of categories as more categories usually lead to a smaller count in each category. On the contrary, UNS's RMSE rises because the magnification of errors during reconstruction is proportional to the number of categories. GNS's accuracy level surpasses UNS when there are 10 or more categories. When there are less than 10 categories, the difference of counts between adjacent categories is relatively significant in GNS. We also observe that the query results of GNS are constantly more accurate than URRP, whose relative accuracy is never higher than 12.3%.

6.1.2. Standard Deviation of GNSs

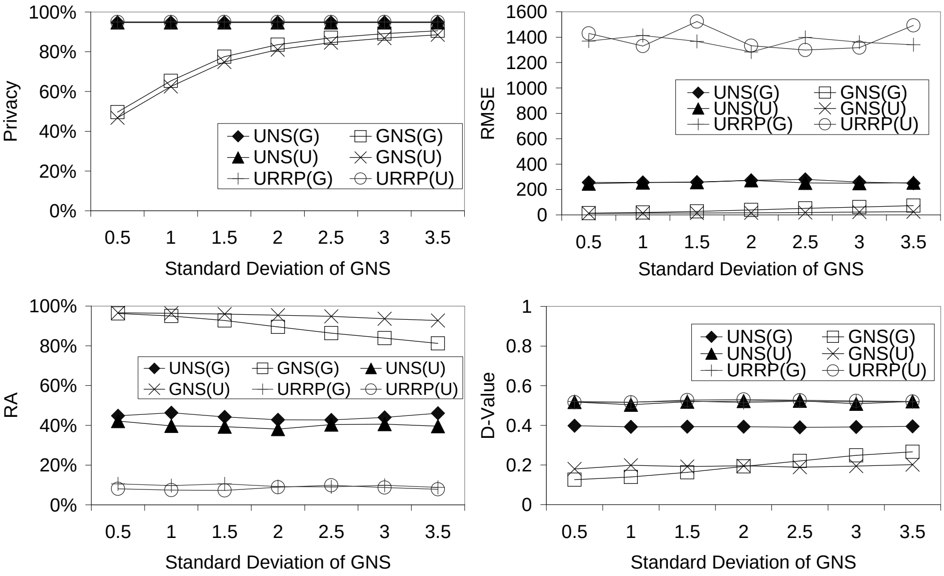

Figure 8: Effects of the standard deviation of GNSs in synthetic one dimensional settings.

The shape of Gaussian distribution used by GNS is determined by the standard deviation. A small value of standard deviation leads to a narrow distribution around the positive category, and vice versa. As expected, the privacy level of GNS is affected by this parameter (Figure 8). When the standard deviation increases, participants select negative categories from a wider range, which improves the privacy levels. GNS's relative accuracy decreases slowly when the standard deviation increases as the true counts become less likely to be distributed near the positive category. However, the results demonstrate that we can increase GNS's privacy level to 90.6% while maintaining accuracy level above 81.2%. The common data perturbation technique, URRP, achieves high privacy levels similar to UNS but are not usable due to poor accuracy levels in solving queries. Based on D-values, we also observe that GNS can approximate the original distribution well, no matter it is Gaussian or uniform. URRP and UNS cannot give close approximation of the original distribution, regardless of the shape of the distribution.

6.1.3. Query Size

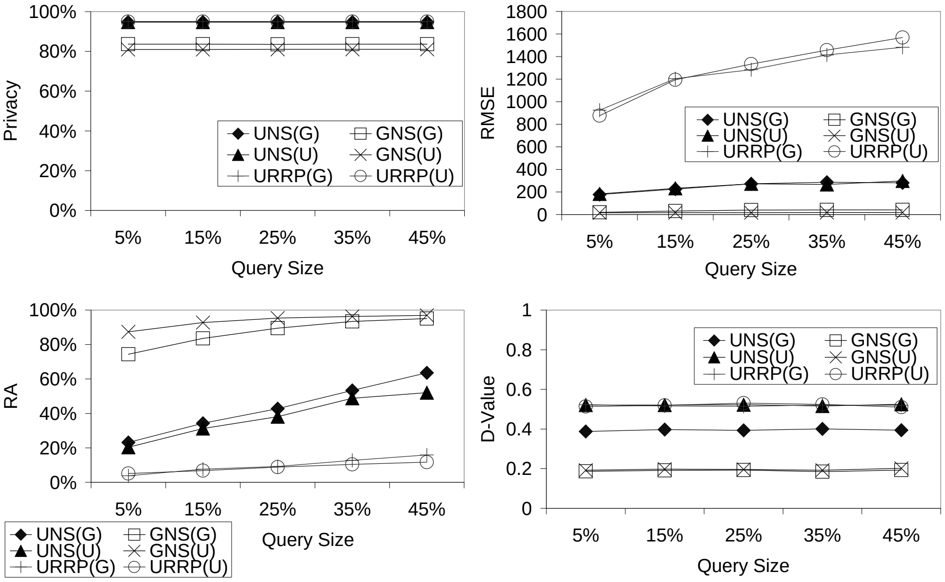

Figure 9: Effects of the query size in synthetic one dimensional settings.

Figure 9 shows that GNS's RMSE is lower than UNS and URRP. GNS also achieves a significantly higher RA than both other approaches. The RA of all approaches increases with the growth of query size. For example, GNS's RA increases from 87.3% to 96.9% when the original data is uniformly distributed. This indicates that the offset of errors from adjacent categories has more significant effects on the results when more categories are involved. Again, the results show that GNS is the only approach that can achieve high privacy levels and high accuracy levels at the same time.

6.1.4. Number of Participants

For all approaches, RMSE increases when there are more participants (Figure 10). GNS's RMSE grows at the slowest pace among all approaches. The relative accuracy of UNS and URRP increases with the growth of participants because the errors become less significant compared to large counts in query results. GNS's relative accuracy is higher than UNS and URRP by more than 37% and 76%, respectively.

6.2. Synthetic Two Dimensional Datasets

Our two dimensional settings simulate data collections in which the number of categories can be quite high. We use the Network-based Generator of Moving Objects [53] to randomly create locations of participants in a part of the road network of the State of Victoria, Australia. We control parameters of the generator such that the number of participants is high in an area where the density of road network is also high, and vice versa. This leads to a realistic distribution of participants. In real data collections, the locations could be home addresses, vehicle positions, etc. We then divide the geographical area into an $n \times n$ grid. Each grid cell is regarded as a category. In UNSs, a

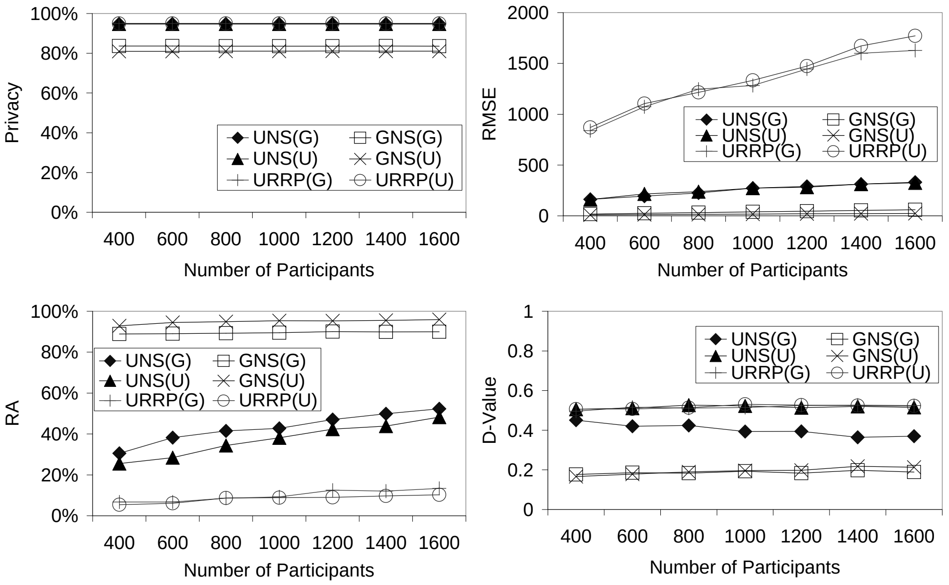

Figure 10: Effects of the number of participants in synthetic one dimensional settings.

participant randomly selects a negative category with a probability of $\frac{1}{n^2-1}$. In GNSs, the probability of selecting a negative category is calculated in the same way as in one dimensional settings, but is based on the number of hops between a true cell and a selected cell. For example, a positive category (a grid cell) may be surrounded by up to 8 one-hop negative categories, each of which has an equal probability to be selected. The larger the number of hops, the lower the probability. We also implement URRPs in this experiment. We define the range of a query as a rectangular region, which covers one or more cells.

6.2.1. Number of Categories

Figure 11 shows the high privacy levels achieved by GNS. The accuracy levels of GNS are significantly better than UNS and URRP. This can be observed from the charts for RMSE, RA and D-value. For example, GNS's relative accuracy is between 87.3% and 94.8% while the other two stay below 10% in most cases. Results on relative accuracy show that UNS are unsuitable for large scale data collections if there are a high number of categories and a limited number of participants. URRP also performs poorly in such situations.

6.2.2. Standard Deviation of GNSs

Figure 12 shows that GNS's privacy level rises with the increment of standard deviation as participants can select negative categories from a wider range. Even when the standard deviation is low, GNS still achieves a 86.7% privacy level. Compared with the results for one dimensional settings, GNS's privacy level is higher for the same standard deviation because there are a substantially larger number of categories to choose from in two dimensional settings. We also observe the significant advantage of GNS over UNS

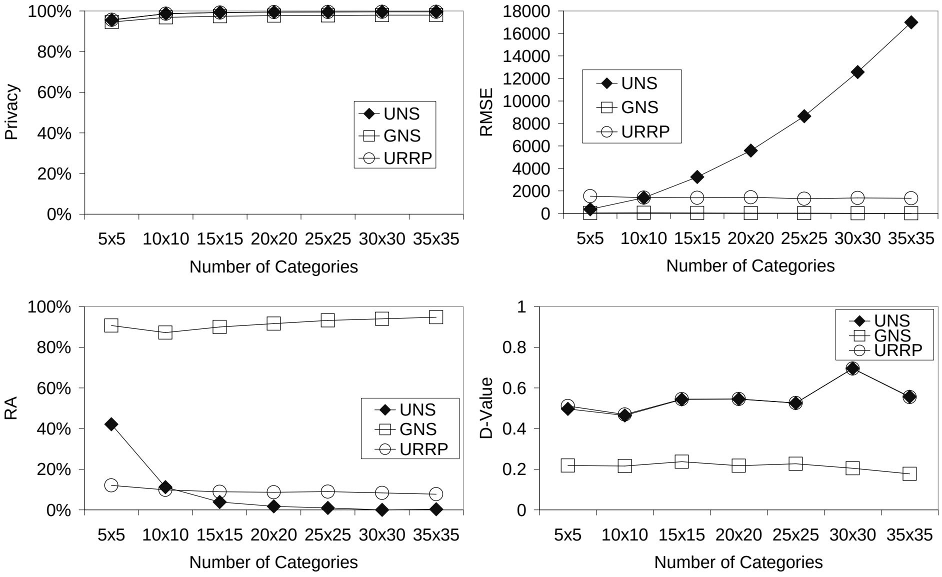

Figure 11: Effects of the number of categories in synthetic 2D settings.

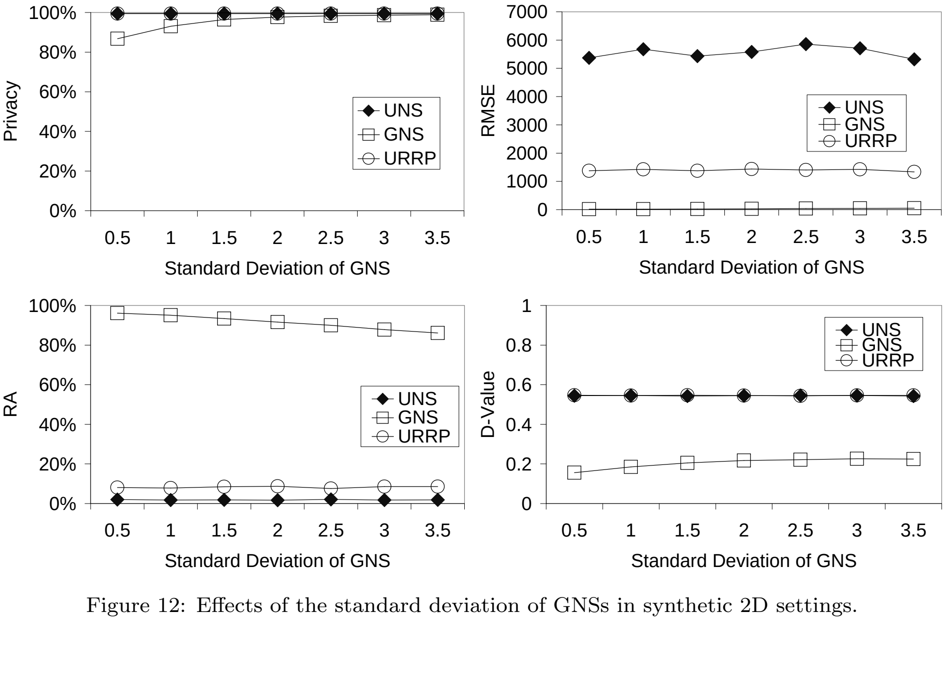

Figure 12: Effects of the standard deviation of GNSs in synthetic 2D settings.

and URRP in terms of accuracy. For example, GNS's D-value is never higher than 0.22 while both UNS and URRP stay at 0.54.

6.2.3. Query Size

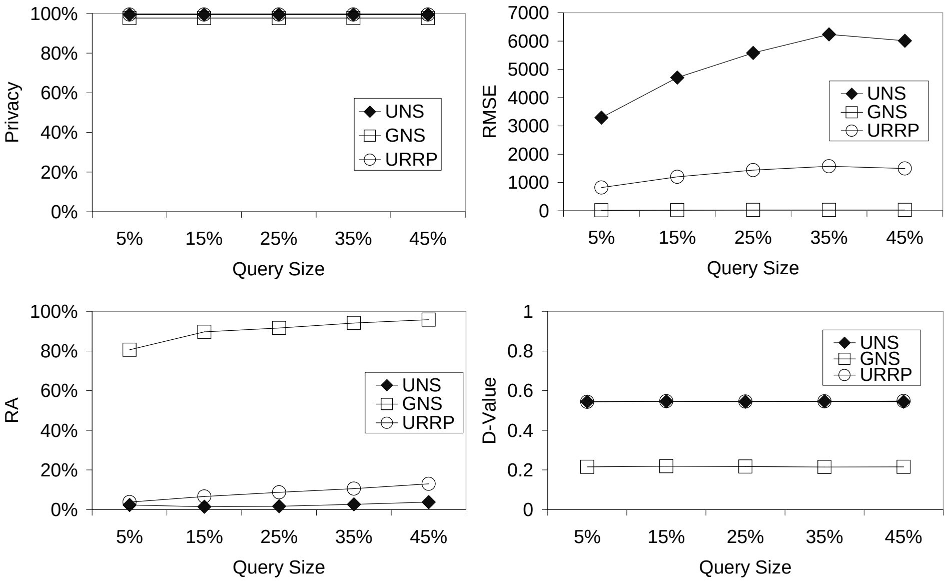

Figure 13: Effects of the query size in synthetic 2D settings.

For any query size, the RMSE of GNS is significantly lower than UNS and URRP (Figure 13). RA of all approaches increases with the growth of query size. GNS's RA is always higher than UNS and URRP by approximately 80%. As shown by the D-values, GNS always shows a more accurate spatial distribution of participants than UNS and URRP.

6.2.4. Number of Participants

Similar to the results from one dimensional settings, Figure 14 shows that GNS is better than other approaches by a large margin in terms of accuracy levels. The number of participants does not affect the high privacy levels of all approaches. Again, the results show that GNS achieves a better balance between privacy and accuracy than UNS and URRP.

6.3. Real Two Dimensional Datasets

In this section, we use a real 2D dataset to test GNSs. We use a dataset that contains the geo-locations of General Practices Surgeries conducted in England during December, 2006 (`http://data.gov.uk/dataset/location_of_general_practices_gps_-_surgeries`). There are 10,201 locations in the dataset. We compute the minimal bounding rectangle of all locations and use each location as the true position of a participant. We then create categories based on a grid in the region as we did in previous

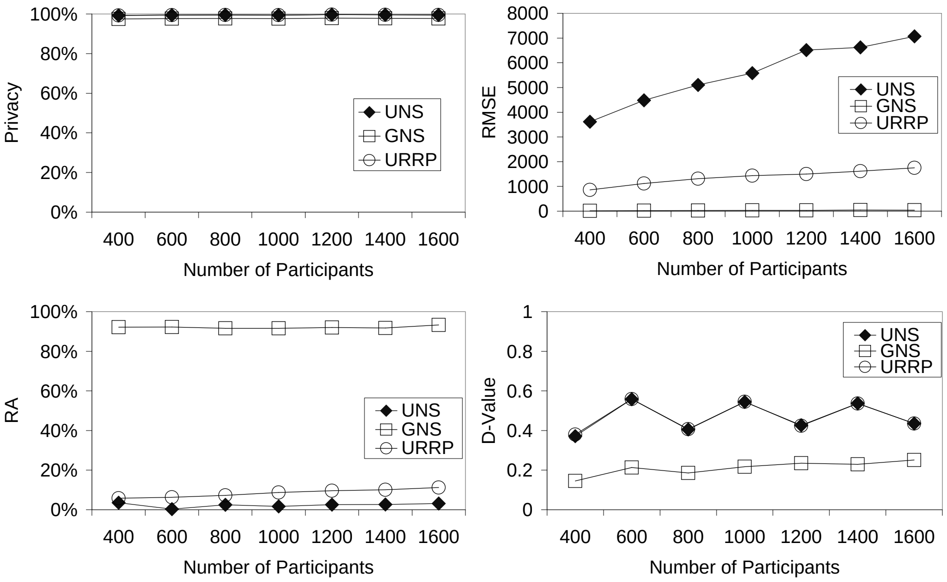

Figure 14: Effects of the number of participants in synthetic 2D settings.

experiments. The true locations of participants are mapped to categories (grid cells). We then generate negative data and compare the answers from the true data and the negative data.

6.3.1. Number of Categories

Figure 15 shows that all three approaches achieve high privacy levels, above 95%. Similar to what we observed in the experiment on synthetic datasets, the accuracy of GNS is better than UNS and URRP in most cases. For example, GNS's RMSE decreases from 2169 to 403 while UNS's RMSE increases from 1155 to 54499. The relative accuracy of GNS is always above 72.7% and increases when there are more categories. At the same time, UNS's relative accuracy drops sharply (82.5% to 3.6%) while URRP stays between (37% and 45%). GNS is the only approach that constantly achieves high privacy levels and high accuracy levels.

6.3.2. Standard Deviation of GNSs

Figure 16 shows that GNS's privacy level catches up with UNS when the standard deviation is 2 or higher. GNS's RMSE is lower by several orders of magnitude than UNS. GNS is also constantly more accurate than URRP in the answers to queries. For example, GNS's relative accuracy ranges from 82.2% to 95.8% while URRP stays around 38%. Based on the D-values, we also observe that GNS shows a more accurate distribution of original data than other approaches when its standard deviation is below 3.

6.3.3. Query Size

When we change the query size, we also observe similar results as in synthetic data (Figure 17). The privacy levels of three approaches are indistinguishable from each other.

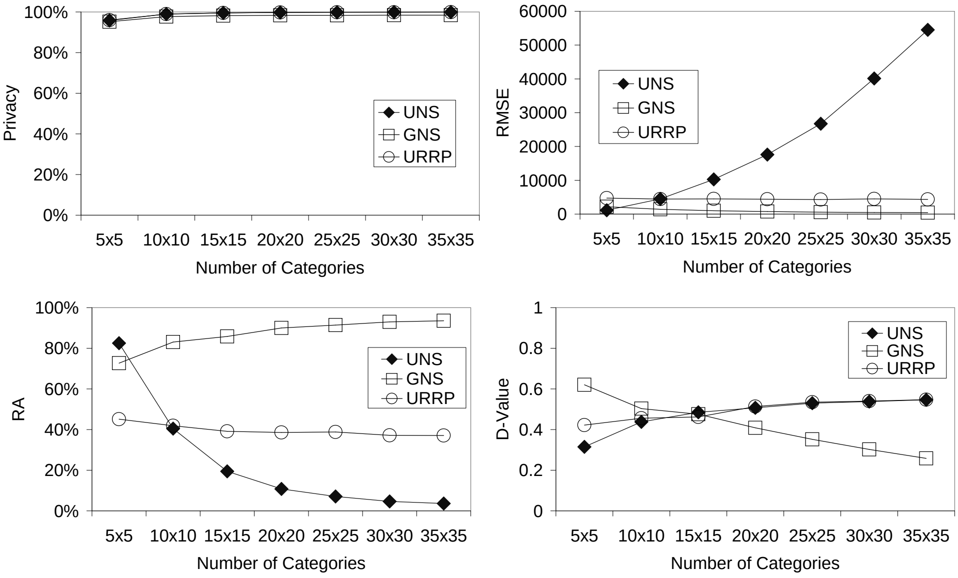

Figure 15: Effects of the number of categories in real 2D settings.

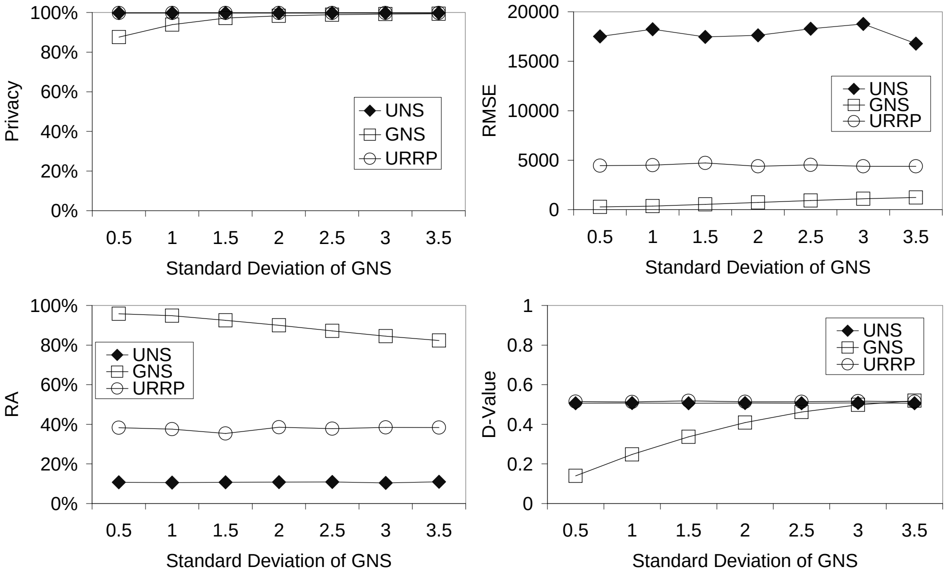

Figure 16: Effects of the standard deviation of GNSs in real 2D settings.

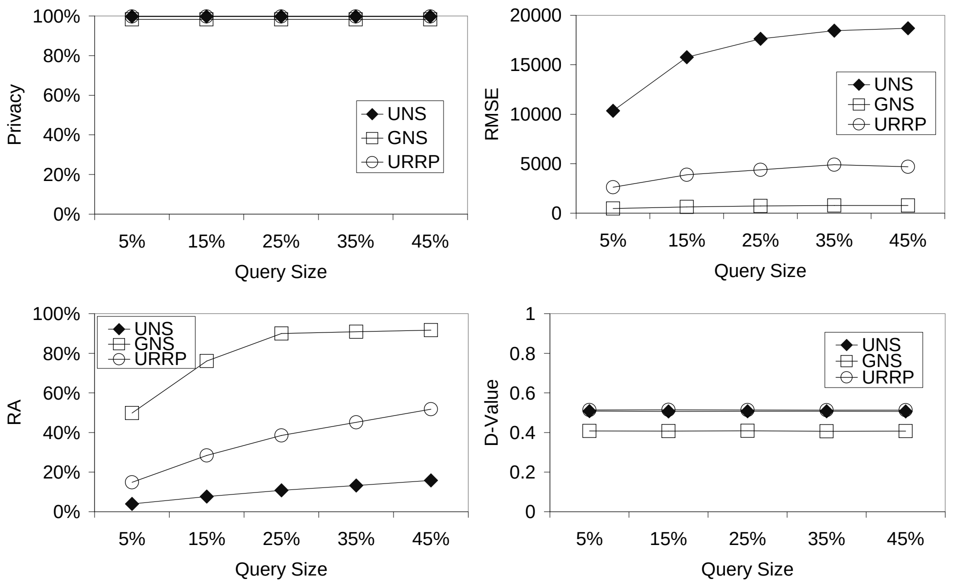

Figure 17: Effects of the query size in real 2D settings.

GNS shows significant advantage in RMSE and relative accuracy over UNS and URRP. For example, GNS achieves 91.6% accuracy level while URRP only reaches 51.8% and UNS is 15.9% when the query size is 45% of the whole area. The D-values of GNS are a bit higher than those in synthetic data due to the high randomness in real data. But they are still constantly lower than UNS and URRP (0.4 versus 0.5). Again, the results show that GNS is the only approach that can achieve high privacy levels and high accuracy levels at the same time.

7. Conclusion

This paper shows that the existing approach for negative surveys, i.e., Uniform Negative Surveys (UNSs), works fine for data collection with a small number of categories but with the increase of the number of categories, the collected data becomes unusable. The effects of the errors are particularly significant when UNSs are used for collecting aggregate spatial data. In this work, we focus on collecting statistical spatial data, although for certain applications, detailed personal data might be more important than such aggregated data. We propose Gaussian Negative Surveys (GNSs). The novelty of GNSs lies in three aspects. First, GNSs maintain the correlation between adjacent categories by using Gaussian distribution rather than uniform distribution as in UNSs. Second, the data collected by GNSs can be used to solve queries without reconstruction, which is mandatory in UNSs. Third, we overcome the limit of survey scale in UNSs, which are only usable when there are a small number of categories. Our experimental results show that GNSs can scale in a significantly better way than UNSs. For example, GNS's RMSE can be lower than UNS's RMSE by several orders of magnitude when there are a number of categories. While UNSs inherently achieve a high privacy level, the privacy level achieved by GNSs is close to UNSs. We also observe that GNSs are significantly more accurate than a common data perturbation technique, URRPs, when both of them achieve similar high privacy levels. The Gaussian distributions of GNSs used in our experiments are multi-variation Gaussian distributions in one dimensional and two dimensional settings. The aim of future study is to apply the same approach to spatial data with a higher number of dimensions, e.g., 3D geo-spatial data or 2D data with another temporal dimension.

References

[1] R. Tourangeau, T. W. Smith, Asking sensitive questions: The impact of data collection mode, question format, and question context, Public Opinion Quarterly 60 (1996) 275–304.

[2] R. Tourangeau, Survey research and societal change, Annual Review of Psychology 55 (2004) 775–801.

[3] M. Duckham, L. Kulik, Location privacy and location-aware computing, in: Dynamic & Mobile GIS: Investigating Change in Space and Time, CRC Press, 2006, pp. 35–51.

[4] L. J. Morgan, Attitudes toward spatial privacy in the United States of America, 2007. Available from: http://gradworks.umi.com/32/61/3261994.html.

[5] R. Hsieh, C. Tutzauer, The influence of privacy and security on respondents' trust and participation in e-surveys, 2003. Available from: http://www.allacademic.com/meta/p111854_index.html.

[6] M. J. Culnan, P. K. Armstrong, Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation, Organization Science 10 (1999) 104–115.

[7] S. Wang, H. Wang, Conceptual construction on incomplete survey data, Data & Knowledge Engineering 49 (2004) 311 – 323.

[8] J. Horey, M. M. Groat, S. Forrest, F. Esponda, Anonymous data collection in sensor networks, MobiQuitous (2007) 1–8.

[9] F. Esponda, Hiding a needle in a haystack using negative databases, Information Hiding: 10th International Workshop (2008) 15–29.

[10] F. Esponda, E. S. Ackley, P. Helman, H. Jia, S. Forrest, Protecting data privacy through hard-to-reverse negative databases, International Journal of Information Security 6 (2007) 403–415.

[11] F. Esponda, S. Forrest, P. Helman, Negative representations of information, International Journal of Information Security 8 (2009) 331–345.

[12] R. Agrawal, R. Srikant, D. Thomas, Privacy preserving OLAP, SIGMOD (2005) 251–262.

[13] N. R. Adam, J. C. Worthmann, Security-control methods for statistical databases: a comparative study, ACM Computing Surveys 21 (1989) 515–556.

[14] B. C. Fung, K. Wang, L. Wang, P. C. Hung, Privacy-preserving data publishing for cluster analysis, Data & Knowledge Engineering 68 (2009) 552 – 575.

[15] W. Yang, S. Huang, Data privacy protection in multi-party clustering, Data & Knowledge Engineering 67 (2008) 185 – 199.

[16] R. Agrawal, R. Srikant, Privacy-preserving data mining, SIGMOD Record 29 (2000) 439–450.

[17] E. Magkos, M. Maragoudakis, V. Chrissikopoulos, S. Gritzalis, Accurate and large-scale privacy-preserving data mining using the election paradigm, Data & Knowledge Engineering 68 (2009) 1224–1236.

[18] S. P. Reiss, Practical data-swapping: the first steps, ACM Transactions on Database Systems 9 (1984) 20–37.

[19] S. E. Fienberg, J. McIntyre, Data swapping: Variations on a theme by Dalenius and Reiss, Privacy in Statistical Databases 3050 (2004) 14–29.

[20] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables, ICDE (2007) 116–125.

[21] Y. He, J. F. Naughton, Anonymization of set-valued data via top-down, local generalization, PVLDB 2 (2009) 934–945.

[22] M. Terrovitis, N. Mamoulis, P. Kalnis, Privacy-preserving anonymization of set-valued data, PVLDB 1 (2008) 115–125.

[23] T. Ge, S. B. Zdonik, Answering aggregation queries in a secure system model, VLDB (2007) 519–530.

[24] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, Order preserving encryption for numeric data, SIGMOD (2004) 563–574.

[25] D. Papadias, P. Kalnis, J. Zhang, Y. Tao, Efficient OLAP operations in spatial data warehouses, Advances in Spatial and Temporal Databases (2001) 443–459.

[26] S. Govindarajan, P. K. Agarwal, L. Arge, CRB-tree: an efficient indexing scheme for range-aggregate queries, ICDT (2003) 143–157.

[27] Y. Tao, D. Papadias, Range aggregate processing in spatial databases, IEEE Transactions on Knowledge and Data Engineering 16 (2004) 1555–1570.

[28] S. Acharya, V. Poosala, S. Ramaswamy, Selectivity estimation in spatial databases, SIGMOD Record 28 (1999) 13–24.

[29] J. Jin, N. An, A. Sivasubramaniam, Analyzing range queries on spatial data, ICDE (2000) 525–534.

[30] Y.-J. Choi, C.-W. Chung, Selectivity estimation for spatio-temporal queries to moving objects, in: SIGMOD, pp. 440–451.

[31] H. Xie, E. Tanin, L. Kulik, Distributed histograms for processing aggregate data from moving objects, MDM (2007) 152–157.

[32] H. Xie, L. Kulik, E. Tanin, Privacy-aware traffic monitoring, IEEE Transactions on Intelligent Transportation Systems 11 (2010) 61–70.

[33] R. Beigel, E. Tanin, The geometry of browsing, LATIN (1998) 331–340.

[34] C. Sun, N. Bandi, D. Agrawal, A. E. Abbadi, Exploring spatial datasets with histograms, Distributed and Parallel Databases 20 (2006) 57–88.

[35] K. Shilton, Four billion little brothers?: privacy, mobile phones, and ubiquitous data collection, Communications of the ACM 52 (2009) 48–53.

[36] K. L. Olson, S. J. Grannis, K. D. Mandl, Privacy protection versus cluster detection in spatial epidemiology, American Journal of Public Health 96 (2006) 2002–2008.

[37] S. Papadopoulos, S. Bakiras, D. Papadias, Nearest neighbor search with strong location privacy, Proceedings of the VLDB Endowment 3 (2010) 619–629.

[38] H. J. Onsrud, J. P. Johnson, X. Lopez, Protecting personal privacy in using geographic information systems, Photogrammetric Engineering and Remote Sensing 60 (1994) 1083–1095.

[39] A. K. W. Yeung, G. B. Hall, User education and legal issues of spatial database systems, Spatial Database Systems (2007) 219–260.

[40] H. Xu, H.-H. Teo, B. Tan, R. Agarwal, The role of push-pull technology in privacy calculus: The case of location-based services, Journal of Management Information Systems 26 (2009) 135–174.

[41] M. Gruteser, D. Grunwald, Anonymous usage of location-based services through spatial and temporal cloaking, MobiSys (2003) 31–42.

[42] M. F. Mokbel, C.-Y. Chow, W. G. Aref, The new Casper: query processing for location services without compromising privacy, VLDB (2006) 763–774.

[43] M. Duckham, L. Kulik, A formal model of obfuscation and negotiation for location privacy, Pervasive Computing 3468/2005 (2005) 152–170.

[44] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. di Vimercati, P. Samarati, Location privacy protection through obfuscation-based techniques, DBSec (2007) 47–60.

[45] S. Wieland, C. Cassa, K. Mandl, B. Berger, Revealing the spatial distribution of a disease while preserving privacy, Proceedings of the National Academy of Sciences 105 (2008) 17608–17613.

[46] M. P. Armstrong, G. Rushton, D. L. Zimmerman, Geographically masking health data to preserve confidentiality, Statistics in Medicine 18 (1999) 497–525.

[47] E. Kaplan, T. B. Pedersen, E. Savas, Y. SaygIn, Discovering private trajectories using background information, Data & Knowledge Engineering 69 (2010) 723 – 736.

[48] S. Mukherjee, M. Banerjee, Z. Chen, A. Gangopadhyay, A privacy preserving technique for distance-based classification with worst case privacy guarantees, Data & Knowledge Engineering 66 (2008) 264–288.

[49] W. R. Tobler, A computer movie simulating urban growth in the detroit region, Economic Geography 46 (1970) 234–240.

[50] L. Sweeney, K-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002) 557–570.

[51] J. Frank J. Massey, The Kolmogorov-Smirnov test for goodness of fit, Journal of the American Statistical Association 46 (1951) 68–78.

[52] G. K. Kanji, 100 Statistical Tests, SAGE Publications London, 2006.

[53] T. Brinkhoff, A framework for generating network-based moving objects, Geoinformatica 6 (2002) 153–180.