Differential Private Trajectory Protection of Moving Objects

Roland Assam
RWTH Aachen University
Germany
assam@cs.rwth-aachen.de

Marwan Hassani
RWTH Aachen University
Germany
hassani@cs.rwth-aachen.de

Thomas Seidl
RWTH Aachen University
Germany
seidl@cs.rwth-aachen.de

ABSTRACT

Location privacy and security of spatio-temporal data has come under high scrutiny in the past years. This has rekindled enormous research interest. So far, most of the research studies that attempt to address location privacy are based on the k-Anonymity privacy paradigm. In this paper, we propose a novel technique to ensure location privacy in stream and non-stream mobility data using differential privacy. We portray incoming stream or non-stream mobility data emanating from GPS-enabled devices as a differential privacy problem and rigorously define a spatio-temporal sensitivity function for a trajectory metric space. Privacy is achieved through path perturbation in both the space and time domain. In addition, we introduce a new notion of Nearest Neighbor Anchor Resource to add more contextual meaning in the face of uncertainty to the perturbed trajectory path. Unlike k-Anonymity techniques that require more mobile objects to achieve strong anonymity; we show that our approach provides stronger privacy even for a single moving mobile object, outliers or mobile objects in sparsely populated regions.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Algorithm, Security

Keywords

Differential Privacy, Location Privacy, Stream Privacy, Moving Object Privacy

1. INTRODUCTION

Although the mainstream adoption of GPS and RFID technologies are invaluable and indispensable in our day to day lives or businesses, the amount of mobility pattern data collected, assembled and analyzed from such services or technologies is eye-opening. The mass storage of mobility data into Moving Object Databases (MOD) or other systems has numerous uses in a broad spectrum of industries. Yet this same data has the potentials to unlock human mobility patterns, human behaviors and other sensitive information.

Gartner Research and its Research VP William Clark[1] are heralding Context Aware Computing as the future of computing. Some very novel and extremely good existing privacy solutions [16], [1], [26] did not take into consideration the context of the location when anonymizing or obfuscation data. In fact, most of the existing research techniques are based on the k-Anonymity [30] and l-Diversity [21] privacy definitions. The aforementioned privacy definitions require at least two mobile objects to achieve anonymity. In addition, they are still prone or exposed to background knowledge attacks, compositional [13] and other attacks. Privacy, especially in terms of location or mobility can not afford to trail behind in context aware computing, which is seen as the worst threat to privacy. As a result of this, in this paper, we employ a privacy paradigm called differential privacy [10] and introduce a new notion of Nearest Neighbor Anchor Resource to add more semantic location context to the differential private results. Providing privacy with a *very strong* privacy paradigm like *differential privacy* in context aware applications will be beneficial to a broad spectrum of fields such as mobile social networking, health service patients surveillance, mobile telecommunications, national security, search engines, traffic monitoring, trend analysis and customer data mining. To the best of our knowledge, this is the first location privacy work that utilizes differential private results to provides a context aware location privacy.

Motivation Example: This paper focuses only on location privacy of moving objects using GPS technology. Let's assume Ann is at a location where she is the only person sending a request to a MOD or Location Based Services(LBS) at a given time. We should note that this happens very often. Although k-Anonymity has provided lots of solutions to tackle location privacy, it will fail to protect Ann. How can one protect the true location of an outlier request from Ann or any individual in a sparsely populated area using a strong privacy definition? This is the first motivation of this paper. However, if there are multiple users including Ann that send requests to the LBS and MOD, k-

ACM SIGSPATIAL IWGS '12, November 6, 2012. Redondo Beach, CA, USA

[1] http://www.gartner.com/technology/research/context-aware-computing/

Anonymity will successfully protect her data. The second motivation is to use an alternative and far stronger privacy paradigm (other than k-Anonymity and which is resistant to background knowledge) called differential privacy to protect multiple users. In both cases, differential privacy is achieved in a context and non-context aware manner.

As a summary, this paper has two main goals. These include the use of differential privacy to ensure non-context aware privacy and secondly, the utilization of differential private outputs to ensure context aware location privacy for outliers or multiple moving objects. We should note that the second goal is not part of differential privacy. The main rational of the second goal is to achieve very good data utility, thus ensuring a trade-off between privacy and data utility.

1.1 Our Contributions

Many research studies dealing with LBS or Trajectory privacy use k-Anonymity [1], [31], [26], [33], [19] or path obfuscation [9], [2], [15] to achieve privacy. Unlike previous works, we use differential privacy instead of k-Anonymity to guarantee the privacy of moving objects. Although the theoretical strength of differential privacy has been highly praised, it is quite difficult and challenging to practically apply it in different domains as mentioned in [29] and [34], partly due to the problems that might be encountered during the derivation of the sensitivity of a metric space. In this paper, we address this challenge, and provide a differential private solution for location privacy using the Laplace noise. In addition, we present a technique that accomplishes context aware location privacy by using differential private outputs for moving objects. Specifically, we propose a novel technique to enforce differential privacy in trajectory data stream by perturbing traces of a given trajectory using differential private noise before sending them to a MOD or LBS. To achieve differential privacy, we first determine a more accurate measurement of an object's true GPS position by partitioning incoming raw GPS data stream readings into different data blocks called Running Window over a constant time slot and then compute their moving averages. Perturbing traces directly is a naive approach which will lead to the addition of too much noise. Hence, we tediously derive the sensitivity of a trajectory metric space and its bounds. Laplace noise drawn from this sensitivity is then added to each Running Window to achieve differential privacy. Furthermore, we introduce a new notion of Nearest Neighbor Anchor Resource (NNAR) in order to add more contextual meaning to a differential private noisy trace. Here is a summary of our contributions:

- we rigorously derive the sensitivity of a trajectory metric space and its bounds by introducing notions like Running Window to capture changes in the metric space.
- we propose a novel technique to achieve differential privacy for spatio-temporal trajectory data streams.
- we present a new notion of Nearest Neighbor Anchor Resource (NNAR) which is needed to achieve context aware location privacy.
- Using real and synthetic datasets, we show that our technique outperforms state-of-the-art previous works.

Paper Organization: The rest of this paper is organized as follows. Section 1.2 focuses on relevant related works. In Section 2, some basic concepts of differential privacy are explained. Section 3 discusses how Laplace noise is used in a trajectory metric space to guarantee differential privacy. In section 4, the notion of NNAR is introduced and utilized to accomplish context aware location privacy. Section 5 discuses the experimental results. In section 6, a final conclusion is made.

1.2 Related Works

Trajectory Anonymization and Location Privacy Techniques such as [16], [25], [14] use the spatial k-Anonymity paradigm. The topography of this paradigm typically comprises of users who send their request through a trusted server to the LBS. Anonymization is accomplished in the trusted server. This is done by selecting an area called cloaking region (CR) and for a given object's request, it ensures that at least *k-1* other object requests in that CR are sent to the LBS. Our approach is similar to this technique only from the setup point of view; the trusted server of the former guarantees privacy through anonymization while our trusted server provides privacy through trace perturbation.

While another technique called SpaceTwist [35] uses anchor location, our NNAR sharply differs from theirs. Unlike SpaceTwist that comprises of demand and supply spaces which constantly move and converge to one another from the start phase to the final phase, our NNAR is basically used to add context to a differential private perturbed trace. k-Anonymity was achieved in [31] by suppression. [31] computes the probability of an adversary to correctly determine a trajectory sequence. It then suppresses certain traces of different trajectories based on these probabilities in such a way that at least k-1 sequences are indistinguishable. [1] and [26] achieved k-Anonymity by generalization. The authors of [1] used inherent GPS error to propose a (k, δ)-Anonymity algorithm called Never Walk Alone (NWA) where δ represents the error radius. They achieved anonymity through space-translation in the process of co-location. Co-location is done by grouping all trajectory traces within δ inside a cylindrical tube. It is worth mentioning that while our technique and [1] deal with GPS, our approach completely differs from [1] in two ways. First, our approach employs the notion of moving average and secondly, we utilize the differential privacy paradigm while the approach in [1] is based on (k-δ)-Anonymity.

Differential Privacy: Fundamental theories of differential privacy are provided in [10], [23], [6] and [5]. We also employ some important guidelines and theories from [11], [27] to derive a sensitivity function for the trajectory metric space which is pivotal in the derivation of a differential private noise.

The data access interface of PINQ [22] and [12] are used for interactive data publishing, while ours and [24] are geared towards non-interactive publishing. PINQ outputs private results for several data mining tasks while [12] is tailored for just the ID3 algorithm. [24] presented a differential private classification technique to generalize data. [20] utilized differential privacy to track commuters' pattern. Differential private frequent item and times-series techniques were proposed by [4] and [29], respectively. [8] propose a differential private spatial decomposition technique which can be utilized to keep GPS traces private. It also provides a differential private version of quadtrees, kd-trees and Hilbert R-trees. Our technique differs from [8] in that we perturbed GPS traces with Laplace noise within a Running Window while [8] presents a novel approach to minimize query error by configuring hierarchical noise parameters in a non-uniform manner.

2. BACKGROUND

2.1 System Setup

There are two ways to publish data privately. These include the *non-interactive* approach and the *interactive* approach. In non-interactive data publishing, the data is first anonymized and then published, so that any data miner can have a copy of the published anonymized data. While during interactive data publishing, the owner of the data keeps the raw data, and data miners that intend to access the raw data must pass through a private data interface layer. This work employs non-interactive data publishing.

The setup consists of a single user or multiple users carrying a GPS- enabled device. As the user(s) or object(s) move, their current spatio-temporal location data are being perturbed and sent to the MOD or LBS via a randomization mechanism as depicted in Figure 1. The randomization mechanism injects a properly calibrated meaningful amount of differential private noise drawn from a trajectory sensitivity function as described in section 3 to create differential private spatio-temporal data. Perturbation of these traces results in the sanitization of an entire trajectory path such that an attacker looking at the perturbed trajectory path in the MOB and LBS cannot determine the original trajectory path.

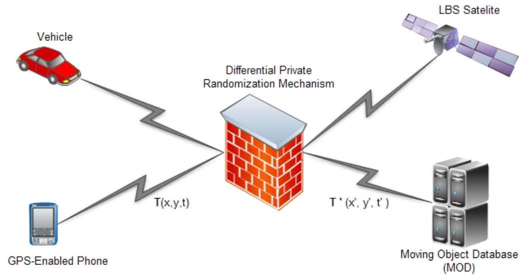


Figure 1: Differential private data interface.

2.2 Basics of Differential Privacy

Differential privacy is a privacy paradigm proposed by Dwork [10] that ensures privacy through data perturbation. Differential privacy is based on the core principle that for any two datasets that differ in only one entry, the ratio of the probability of the outputs of their randomized computations is very small.

Example: Consider a dataset $\mathcal{T}_1$ that has 50 records. If one record is added or removed from the dataset $\mathcal{T}_1$, a new dataset $\mathcal{T}_2$ which contains 51 or 49 records is formed, respectively. If a query is separately run on $\mathcal{T}_1$ and $\mathcal{T}_2$, and a calibrated amount of noise is added to the true query results by a randomized mechanism; if the randomized mechanized obeys differential privacy, the ratio of the probability of the output query results of both datasets $\mathcal{T}_1$ and $\mathcal{T}_2$ will be small. This means, the presence or absence of a single record from a dataset does not leak any information, thus providing privacy. This is formally given as follows.

Definition 1. (ϵ-differential privacy [11]): *A randomization mechanism $\mathcal{A}$ (x) provides ϵ-differential privacy if for any two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ that differ on at most one element, and all output $\mathcal{S} \subseteq$ Range($\mathcal{A}$),*

$$Pr[\mathcal{A}(\mathcal{D}_1) \in S] \leq \exp(\epsilon) * Pr[\mathcal{A}(\mathcal{D}_2) \in \mathcal{S}]$$

The above definition simply means, the randomization process ensures that regardless if an individual chooses to include or remove her record from a dataset, there would be negligible change at the output, thus guaranteeing privacy. ϵ is the privacy parameter called privacy budget or privacy level. When ϵ is less than one, then $\exp(\epsilon) \approx 1 + \epsilon$

Sensitivity: In differential privacy, *sensitivity* is very critical during the process of noise derivation. For a dataset consisting of several inputs, the sensitivity is defined as the maximum change that occurs if one input is removed from the dataset. Formally, the sensitivity is defined as follows.

Definition 2. ($\mathcal{L}_1$ SENSITIVITY [11]): *The $\mathcal{L}_1$ sensitivity of a function $f : D^n \to \mathbb{R}^d$ is the smallest number S(f) such that for all x and x' which differ in a single entry,*

$$\|f(x) - f(x')\| \leq S(f)$$

In this paper, we define a new notion of sensitivity in Section 3.4 which is specifically tailored for a trajectory metric space.

Noise Addition: Differential privacy is achieved by adding noise to data. Three types of noise can be used. These include, the Laplace noise, the Gaussian noise and the Exponential Mechanism [23]. This study uses the Laplace noise to achieve differential privacy.

Laplace Noise: The Laplace noise [11] is drawn from the probability density function of the Laplace distribution. The Laplace noise is said to be ϵ-differential private if Theorem 1 is satisfied.

Theorem 1. *For a given function $f : D^n \to \mathbb{R}^d$, which has sensitivity S(f) , a mechanism $A(x) = f(x) + Lap(\frac{S(f)}{\epsilon})^d$ provides ϵ-differential privacy.*

Composition: [22] mentioned that there are basically two types of compositions. These include, *Sequential Composition* and *Parallel Composition*. Sequential composition is exhibited when a sequence of computations provides differential privacy in isolation. The final privacy guarantee is said to be the sum of each ϵ-differential privacy. On the other hand, parallel composition occurs when the input data is partitioned into disjoint sets, independent of the original data. In this case, the final privacy from such a sequence of computation depends on the worst computation guarantee of the sequence.

3. TRAJECTORY DIFFERENTIAL PRIVACY

Foundational theoretical works [10], [11] of differential privacy explained how to achieve differential privacy for predicate outputs (i.e. 0 or 1). However, many real life applications have more complex outputs. Hence, in order to achieve differential privacy for trajectories, trajectory data needs to be intensively analyzed, re-defined and modeled to capture changes in a trajectory metric space.

3.1 GPS Precision

GPS position measurements are inaccurate and these inaccuracies might arise from a broad range of factors such as sampling rate, number of available satellites, poor antenna or hardware. This problem is widely documented, e.g. in [3],[28]. Lots of GPS researchers including [32],[18] carried out extensive in-depth investigations and research on how to reduce the errors associated with GPS measurements. Their researches stipulate that the use of *Moving Average Filters* can enhance the precision of GPS measurements. Research works in [7],[17] further support the fact that a GPS position derived from a moving average filter is more accurate than that from a raw GPS reading.

In this paper, we consider GPS positions computed from a moving average filter as the true geographic location of an object instead of directly using raw GPS data, simply because of their high precision. Hence, raw spatio-temporal GPS readings got from GPS satellites which are associated with some errors, will be made accurate using moving average filters. We build such a moving average filter by grouping instances of raw GPS spatio-temporal data of a mobile object into data blocks over a specified time slot. Each data block partition is called a **Running Window**. The time span that bounds each Running Window is constant and the periodicity of this time slot should take into consideration that an adequate amount of data is present in the Running Window.

Definition 3. (Running Window): *is a partitioned data block that comprises of a finite amount of raw GPS spatio-temporal data.*

Creating High Precision Locations: A high precision geographic location of an object is determined from a Running Window by computing the moving average using the raw GPS points within that Running Window.

Definition 4. (High Precision Trace): *A High Precision Trace is a spatio-temporal data whose spatial and temporal values are determined by computing a moving average for each domain using the moving average filter function within a Running Window.*

The moving average filter function $f(x)$ is formalized in Equation 1.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

3.2 Problem Definition

Notations: Let $\mathbf{RTr}_i$ be a set of raw GPS points where $i \in \{1, 2, ...n\}$. A single raw GPS point of $\mathbf{RTr}_i$ is termed a ***Trace***, and each trace given by $(x_i,\ y_i,\ t_i)$ corresponds to a geographic position $(x_i,\ y_i)$ at time t_i. The set of raw GPS traces $\mathbf{RTr}_i$ does not represent the exact geographic positions of an object because of the errors associated with GPS measurements. A more accurate geographic position of an object called the *high precision trace* is determined by partitioning $\mathbf{RTr}_i$ into several Running Windows $\mathbf{W}_j$ where $j \in \{1, 2, 3, ...m\}_{m<n}$ and computing the moving average of $\mathbf{W}_j$. Let $\mathbf{HTr}_j$ denotes a high precision trace of $\mathbf{W}_j$. $\mathbf{HTr}_j$ is considered as the location of the object. Like $\mathbf{RTr}_i$, the high precision trace $\mathbf{HTr}_j$ is a spatio-temporal data given by $(x_j,\ y_j,\ t_j)$.

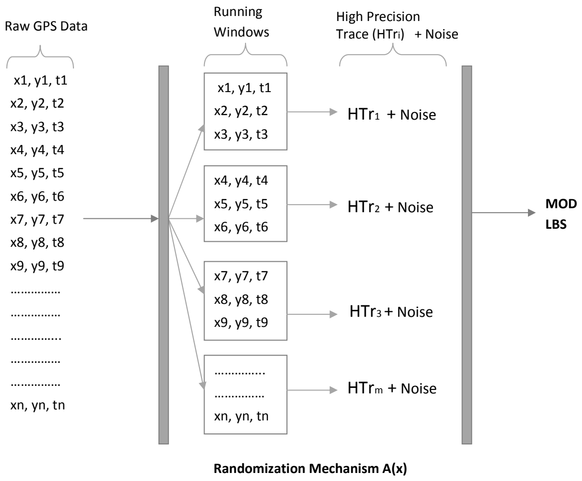


Figure 2: Differential Privacy in each Running Windows.

Definition 5. (Problem Definition): *Assume that an outlier moving object $\mathcal{M}$ sends a stream of raw spatio-temporal GPS data traces $\boldsymbol{RTr}_i$ to a randomized mechanism $\mathcal{A}$ (x). Also consider that the mechanism periodically generates a high precision trace $\boldsymbol{HTr}_j$ of $\mathcal{M}$ from Running Windows $\boldsymbol{W}_j$, and then computes the moving average of a Running Window. Perturb the high precision trace $\boldsymbol{HTr}_j$ by adding differential private noise to it (in both space and time domains) to produce a* ***perturbed*** *trace $\overline{\boldsymbol{HTr}}_j$, such that the ϵ-differential privacy condition is fulfilled. The perturbed trace $\overline{\boldsymbol{HTr}}_j$ should then be sent to the LBS or MOD.*

The problem definition requires that when raw GPS spatio-temporal data stream is sent to the randomization mechanism; the data is first partitioned into Running Windows and a high precision trace is calculated. Then the next and main task of the randomized mechanism is to perturb the high precision trace differential privately and send each domain of the perturbed trace to the MOD or LBS as shown in Figure 2.

Definition 6. (Problem Definition 2): *Given a differential private perturbed high precision trace $\overline{\boldsymbol{HTr}}_j$ and a defined radius $\boldsymbol{r}_{nn}$, determine the nearest neighboring noisy location with the minimum Euclidean distance to $\overline{\boldsymbol{HTr}}_j$ such that:*

- *If the distance between any NNAR location and $\overline{\boldsymbol{HTr}}_j$ is within $\boldsymbol{r}_{nn}$, it will be considered as the noisy location.*
- *Otherwise $\overline{\boldsymbol{HTr}}_j$ is considered as the noisy high precision trace.*

Simply put, after a differential private perturbed trace has been computed, the NNAR algorithm searches for a neighboring location from the NNAR resource pool and checks if the latter location is within a given radius. If it is within the radius, it will be considered as the noisy location. If not, the perturbed high precision trace $\overline{\mathbf{HTr}}_j$ is taken as the noisy location. NNAR is used to add more contextual meaning to the differential private noisy location. Section 4 provides a detail description of NNAR.

3.3 Linking Differential Privacy to Trajectory

Overview

The difficulties of practically implementing differential privacy in other domains were listed in Section 1. This section provides a comprehensive research solution to this challenging problem for a trajectory metric space. In Section 3.1, we alluded that a high precision trace which highlights the true location of an object is computed using a moving average filter. At a higher level, we infuse differential privacy into a trajectory metric space by probabilistically injecting Laplace noise to the moving average filter during the computation of a high precision trace. This prompts the production of differential private noisy moving averages at the output, which correspond to differential private high precision traces. Then, instead of sending the high precision trace to the MOD or LBS, the differential private noisy high precision trace is sent to the MOD or LBS.

Trajectory Perturbation

During the proposal of the Laplace noise in [11] only predicate outputs (i.e. 0 or 1) were considered. Adding Laplace noise in a trajectory metric space is quite challenging because the noise have to be well calibrated with a sensitivity function that captures changes in the trajectory metric space. Our privacy settings is illustrated in Figure 2. Naively adding Laplace noise within each Running Window will produce outputs which have very low utility or noisy outputs whose semantic locations have no meaning.

To address this hurdle, we portray the events in each Running Window as a probabilistic process. Specifically, we consider the dataset $\mathcal{T}_1$ that corresponds to a collection of raw GPS data within a given Running Window as the original dataset. Removing one raw GPS spatio-temporal data from that Running Window forms a new dataset $\mathcal{T}_2$ such that $\mathcal{T}_1$ and $\mathcal{T}_2$ differ in just one single entry. Then during the computation of a moving average in that Running Window, the randomize mechanism $\mathcal{A}(x)$ adds a carefully calibrated amount of Laplace noise to the true value of the moving average to form a noisy moving average with a probability derived from the randomness of $\mathcal{A}$. The Laplace noise is drawn from a sensitivity function derived in the next section (Section 3.4).

3.4 Sensitivity Function

The definition of a sensitivity function is pivotal for noise derivation. Sensitivity is defined as the maximum possible change that occurs when a single point is removed from a dataset. For a trajectory metric space, we make use of the natural property of the high precision measurement of a GPS geographical location using the moving average. Since the high precision traces are destined to be sent to the MOD or LBS, and are derived from the moving average function, this moving average function will play an important role during the derivation of the sensitivity function of the trajectory metric space. The moving average filter function is given by Equation 1 in Section 3.1.

Sensitivity: To determine the sensitivity, we find the maximum possible change that will occur when one raw GPS spatio-temporal point is removed from the dataset $\mathcal{T}_1$ to form a dataset $\mathcal{T}_2$ within a Running Window. This sensitivity is given by Equation 2.

$$S(f) = \max_{\mathcal{T}_1, \mathcal{T}_2} \|f(\mathcal{T}_1) - f(\mathcal{T}_2)\|_1 \qquad (2)$$

Bounds of the Sensitivity Function: We envision a scenario whereby an object moves and halts at different locations, since this typically reflects human mobility behavior. During our consideration of Laplace noise, we also took this scenario into account. Theoretically, the randomness associated with the addition of Laplace noise is high such that even if a person stays at a given location for a while, different noisy locations will be emitted to the LBS or MOD each time. However, due to data utility concerns, it is important to ensure that the true location of an object and the differential private location of the object is meaningful and can be useful for data mining. As a result of this, we theoretically investigate the bounds of the sensitivity function, since the latter function strongly influences the magnitude of the Laplace noise. The sensitivity function depends on the high precision trace which is got from a Running Window. The lower and upper bounds of any high precision trace is given by Lemma 1. Simply put, Lemma 1 stipulates that the upper bound of any high precision trace (i.e. the result of the moving filter) is less than or equal the position of the highest raw GPS point in that Running Window.

LEMMA 1. *For each domain, the maximum change within a Running Window occurs if the trace with the smallest numerical value is removed.*

$$\max\left(|f(\mathcal{T}_1) - f(\mathcal{T}_2)|\right)_{\forall \mathcal{T}_1, \mathcal{T}_2} \leq \frac{1}{N-1} \sum_{m=1}^{N-1} A_m \qquad (3)$$

$\forall m \neq \min(A)$ *where* A_m *is a finite set* A *consisting of* m *raw GPS point, and* $\min(A)$ *denotes a trace with the minimum magnitude for a given domain in a Running Window.*

See prove in Appendix A. Thus, the bounds of the sensitivity function is finite and not quite large. This will translate to moderate magnitudes of noise and good utility.

3.5 Differential Private Trajectory Algorithm

Noise Addition: Algorithm 1 depicts the differential private perturbation algorithm. The algorithm takes in the following parameters as inputs. The privacy level ϵ, the dimensions d to be perturbed (space and time), the number of raw GPS traces to be allocated in each Running Window $\mathbf{m}$, as well as a non-stream or stream GPS trajectory as its dataset $\mathcal{T}_1$. Streams of incoming raw GPS data are separated into Running Windows (Line 1) based on $\mathbf{m}$ and are used to compute high precision traces in Line 2 by employing the moving average filter function. The sensitivity function which is required for the derivation of Laplace noise is computed in Line 3. In Line 4, the Laplace noise derived from the latter sensitivity function is injected to the result of the moving average of a Running Window to output a differential private noisy moving average. This is performed for each of the selected domains or dimensions specified by d. The moving averages correspond to high precision traces $\mathbf{HTr}_j$, likewise, the noisy differential private moving averages $\overline{\mathbf{HTr}}_j$ are equivalent to differential private high precision traces.

Analysis of Privacy Guarantee: All noisy high precision traces emanating from the trusted server are differentially private.

THEOREM 2. *Algorithm 1 is ϵ-differentially private.*

PROOF. In Line 4 of algorithm 1, Laplace noise is added. Theorem 1 states that the addition of Laplace noise guarantees differential privacy. Also, Line 4 is performed only once for a given Running Window and a given dimension. Since Laplace noise of $Lap\left(\frac{S(f)}{\alpha}\right)$ is used for the perturbation of data in a Running Window, then according to Theorem 1, Line 4 guarantees $1 \times \alpha$-differential privacy. However, because a spatio-temporal data contains three dimensions, namely the X-position, Y-position and the time domain, the privacy budget needs to be carefully managed to control the cost of privacy. Using the Sequential Composition [22] described in Section 2, the total cost of privacy in a Running Window to perturb the different dimensions is $\alpha.|D|$. Where $|D|$ is the number of dimensions and $2 \leq |D| \leq 3$. The set of raw GPS traces from one Running Window do not intersect with raw GPS traces from another Running Window. Since the datasets from a Running Window are independent and disjoint from each other, it implies, following the principle of Parallel Composition [22], the privacy budget does not need to be shared across Running Windows. Hence each Running Window remains $\alpha.|D|$-differential private.

This means, if all domains of a spatio-temporal trace are perturbed (i.e. $|D| = 3$) then each Running Window is 3α-differential private. On the other hand, if only the spatial domains of a trace are perturbed, then each Running Window will be 2α-differential private. Thus, for a given Running Window dataset, each noisy high precision trace sent to the MOD or LBS after noise addition by the Laplace mechanism is $\alpha.|D|$-differential private.

Therefore, if an overall privacy budget ϵ is provided by the data miner, for $\alpha = \frac{\epsilon}{|D|}$, Algorithm 1 is ϵ-differential private. □

Algorithm 1: Differential Private Trace Perturbation

```
Input: Dataset $\mathcal{T}_1$, privacy budget $\epsilon$, number of
       dimensions d, number of raw GPS points per
       Running Window $\boldsymbol{m}$
Output: differential private Noisy High Precision
        Trace ($\overline{\mathbf{HTr}}_j$)
1 Partition: Partition and group $\boldsymbol{m}$ raw GPS points into
    a Running Window
2 Moving Average: Compute a High Precision GPS
    trace ( $\boldsymbol{HTr}_j$ ) from Running Window j using Equation
    1
3 Sensitivity: Get the sensitivity $S(f)$ of the trajectory
    metric space using Equation 2, $\mathcal{T}_1$ and $\mathcal{T}_2$; where $\mathcal{T}_2$ is
    formed by removing a point from $\mathcal{T}_1$ for each Running
    Window
4 Perturbation: Add Laplace noise of $Lap(\frac{S(f)}{\alpha})^d$ to
    the moving average of a Running Window $j$ to
    determine the Noisy High Precision Trace $\overline{\mathbf{HTr}}_j$
5 return: send $\overline{\boldsymbol{HTr}}_j$ to MOD or LBS
```

4. LOCATION DATA UTILITY

This section focuses on context aware location privacy and the novel notion of Nearest Neighbor Anchor Resource (NNAR). It provides a motivational example, explains the core concept and finally presents the NNAR algorithm. NNAR is not a mandatory component of the main differential private algorithm presented in the previous section. It could optionally be used to immensely enhance data utility while preserving privacy.

4.1 Context Aware Location Privacy

Context Aware computing motivations were described in Section 1. The second problem definition (Definition 6) requires the determination of a nearest neighboring location, whose semantic location context is similar to that of the trace that is suppose to be perturbed through differential privacy. We should stress that the notion of *Nearest Neighbor Anchor Resource* and Section 4 as a whole is not part of differential privacy. The main motif of this section is to improve the semantic location context of a differential private high precision trace, thereby ensuring good data utility.

NNAR Motivation Example: Lets assume that the true location of Ann is a McDonald fast food restaurant. Consider that our differential private randomized mechanism creates and sends a noisy location (i.e. noisy high precision trace) to the MOD or LBS which is closer to that McDonald restaurant. The latter noisy location can be a shop or a road. In the worst case scenario, such a noisy location can be a forest, river, lake, waste depot etc. Lets assume that the noisy location is a river. If this noisy location (river) is sent and stored in an MOD or LBS, and seven months later, a data miner intends to analyze this data for trend analysis without hurting Ann's privacy, the miner will conclude that Ann went to (or might like) that river. This is a loss of valuable contextual information. A much better result would have been: Ann went to a restaurant without revealing which restaurant.

However, on one hand, Ann's privacy has been preserved and few knowledge could be gained from the fact that Ann is hanging around the vicinity of the McDonald restaurant. On the other hand, there is loss of important contextual information which could be gained by the data miner without compromising Ann's privacy. Moreover, the result might even become completely meaningless if a trajectory path of Ann switches from road to river to river to road to river. An attacker will notice instantly that such a mobility pattern is not real, and this undermines privacy.

In this work, we introduce and employ the notion of NNAR to add more contextual meaning to noisy differential private traces thereby ensuring better data utility without violating privacy concerns.

NNAR: To some users, knowing the context of their locations is an intrusion to their privacy. In contrast, there are other users who would like to share (some of) the context of their locations which do not hurt their privacies. As a result of this, NNAR is user driven in order to put a user at the driver's sit of her privacy. The user is given the privilege to specify multiple categories that she wishes to emit to the LBS or MOD. The server generates a resource pool based on the user's current location. This is utilized in Algorithm 2 to output a context aware location. The resource pool of a NNAR is basically a list of location coordinates mapped to some categories. These categories are generated by the server, and they have similar location context to the original location. An example of such categories include roads, shops, banks, restaurants. Table 1 illustrates the data structure of a NNAR resource pool. The nature of the user defined categories should provide a high level generalization of the actual location. Sub-categories such as place names and addresses are prohibited as this might disclose private information.

4.2 NNAR Algorithm

NNAR Resource Pool		
Latitude	Longitude	Category
23.845089	38.018470	Road
23.845179	38.018069	Bank
23.845530	38.018241	Shop

Table 1: Nearest Neighbor Anchor Resource Pool

Algorithm 2 depicts of our NNAR algorithm. When a differential private noisy high precision trace is inputed into Algorithm 2, the NNAR algorithm uses the user specified category to search and choose a list of locations in the resource pool which has the same category as the high precision trace (*not the differential private noisy high precision trace*) and has the shortest distance to the differential private noisy high precision trace (Line 3 - 11). This search is basically done by computing the Euclidean distance between the noisy high precision differential private trace and each location in the NNAR resource pool for that category (Line 4). The resource with the minimum Euclidean distance is chosen. The algorithm then verifies if the chosen location is within a given radius $\mathbf{r}_{nn}$.

There are three possible scenarios which might occur during data transmission to the MOD and LBS.

1. If a location from the resource pool is found within $\mathbf{r}_{nn}$ and that resource pool location is not $\mathbf{HTr}_j$, then that location will be sent to the LBS or MOD.

2. If a location from the resource pool is found within $\mathbf{r}_{nn}$ but the resource pool location is equivalent to $\mathbf{HTr}_j$, then the noisy differential private trace $\overline{\mathbf{HTr}}_j$ will be sent to the LBS or MOD.

3. If no location is found from the resource pool, the noisy differential private trace $\overline{\mathbf{HTr}}_j$ will be sent to the LBS or MOD.

As a summary of the above, the differential private noisy high precision trace $\overline{\mathbf{HTr}}_j$ will be replaced by a location from the resource pool only if a NNAR is within the specified radius $\mathbf{r}_{nn}$ and that location is not $\mathbf{HTr}_j$.

Algorithm 2: NNAR Algorithm

```
   Input: High Precision Trace $\mathbf{HTr}_j$, Noisy High
          Precision Trace $\overline{\mathbf{HTr}}_j$, Category $Cat$, Vicinity
          Radius $\mathbf{r}_{nn}$, Resource Pool $LocPool$
   Output: Context Aware candidate traces
1  $locationWithMinDistance \leftarrow \overline{\mathbf{HTr}}_j$ ;
2  $minDistance \leftarrow NULL$;
3  while (categoryOf($\mathbf{HTr}_j$, Cat) in LocPool ) do
4      $currentDistance \leftarrow$ computeEuclideanDistance(
       $\overline{\mathbf{HTr}}_j$, $currentLocation_{pool}$);
5      if ( this is the first LocPool Entry ) then
6          $minDistance \leftarrow currentDistance$ ;
7      end if
8      if ($currentDistance < minDistance$ ) then
9          $minDistance \leftarrow currentDistance$ ;
10     end if
11 end while
12 if (minDistance is within $\mathbf{r}_{nn}$ AND
   locationWithMinDistance is not $\mathbf{HTr}_j$ ) then
13     return: $locationWithMinDistance$;
14 end if
15     return: $\overline{\mathbf{HTr}}_j$ ;
```

5. EMPIRICAL EVALUATION

The implementations were done in Java and the experiments were conducted on an Intel PC with 4GB RAM. To examine the strength and effectiveness of our technique, we based our evaluations on two criteria. 1) Quantifying Privacy obtained by the user. 2) Quality and Utility of the obfuscated trace to databases and data mining. In each of these criteria, we compared our technique with that of two state-of-the-art works. They include the Never Walk Alone (NWA) algorithm [1] and the Path Confusion (PPC) algorithm [15]. Throughout this section, we will refer to these previous works as NWA and PPC, respectively. In the k-Anonymity privacy paradigm, k denotes the number of indistinguishable objects. We should note that throughout this section, our technique which is based on differential privacy does not compare k (from k-Anonymity) to ϵ (from differential privacy). Instead, in order to orchestrate that our technique preserves the privacy of outliers, we compare and highlight from time to time the number of moving objects used in our technique to the value of k used by a k-Anonymity technique.

5.1 Experimental Dataset

We conducted our experiments with one synthetic dataset and two real dataset. The Brinkhoff[2] Oldenburg synthetic dataset was used. We generated 101,070 traces. Besides, we utilized 90,104 traces from the GeoLife [36] Microsoft Asia human mobility real datasets. The Geolife dataset entails the mobility history of 165 users mostly around Beijing and China. In addition, the Athens Truck[3] real dataset that entails 276 GPS trajectories of 50 moving trucks in Athens and a total of 112203 location traces was used.

5.2 Quantifying User's Privacy

We utilized two location privacy metrics to analyze the privacy obtained by a user during perturbation. They include 1) Expectation of Distance Error and Quality of Service (QoS) 2) Location Entropy.

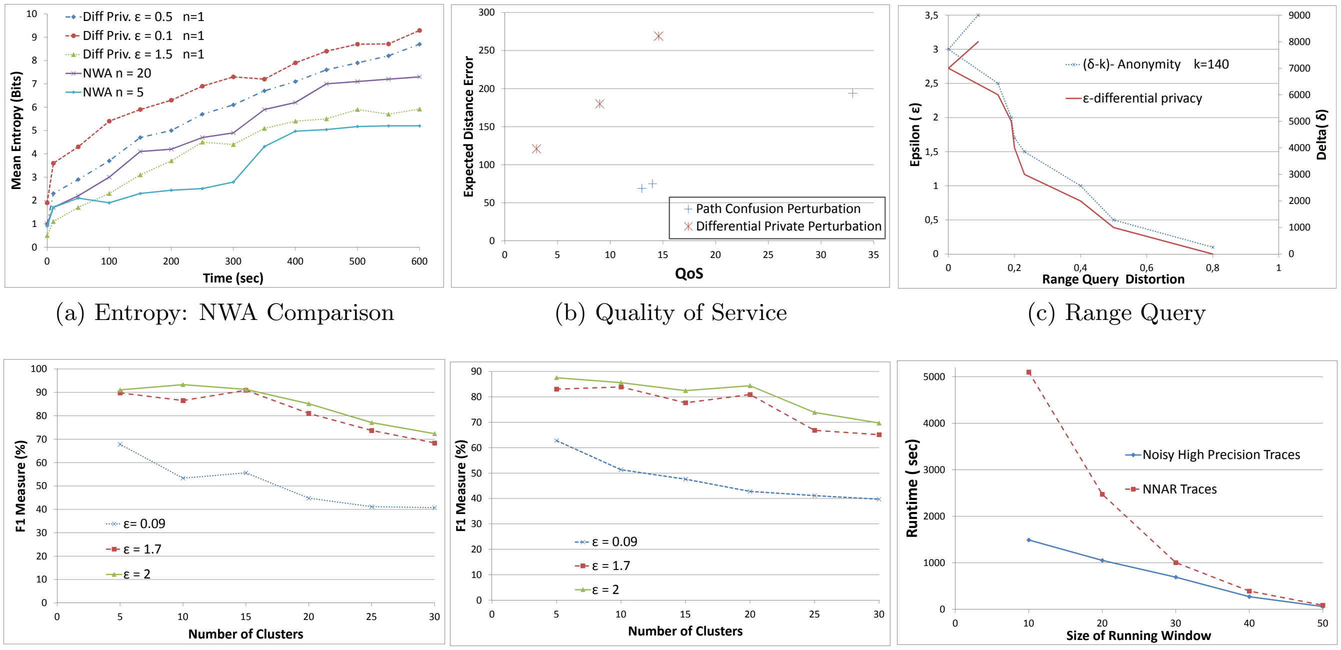
Expectation of Distance Error and QoS: These privacy metrics were proposed by [15]. Expectation of Distance Error measures the accuracy by which an adversary can estimate the true position of a moving object. It is given by:

$$E[d] = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{I} p_i(t)d_i(t) \tag{4}$$

where N is the number of objects, d_i denotes the total distance error between the true and perturbed location, T the total observation time and $p_i(t)$ is the probability to track a user. While [15] used the Reid's Multi-Hypothesis tracking algorithm to determine $p_i(t)$, we customized the definition of $p_i(t)$ to our approach since it makes more sense. We assumed that the adversary has background knowledge of where the user to be protected can roughly be (which is the worst case scenario) and $p_i(t)$ is given by the ability of an

[2] http://iapg.jade-hs.de/personen/brinkhoff/generator/

[3] http://www.rtreeportal.org



(a) Entropy: NWA Comparison

(b) Quality of Service

(c) Range Query

(d) F1 Measure Context Aware

(e) F1 Measure Differential Private Trace

(f) Runtime: Running Window

Figure 3: Evaluation of differential private trace obfuscation.

adversary to predict the correct trace. On the other hand, **QoS** is given by:

$$QoS = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \sqrt{\sum_{j=1}^{J} \left(\widetilde{a_n}(t) - a_n(t)\right)^2} \tag{5}$$

where a is the domain, $a_n(t)$ is the true trace and $\widetilde{a_n}(t)$ the perturbed trace of user n at step t.

We passed the Geolife dataset which has a GPS sampling rate of 2 to 4 seconds into the randomized mechanism. The raw GPS data for each trajectory was partitioned into blocks of 50 traces per Running Window because of their low sampling rate. We considered the movement of a user with the GeoLife dataset and perturbed the traces using our technique. We computed $E[d]$ and QoS, and compared our results with that of PPC. Figure 3b illustrates these results; our technique delivers a better QoS than the PPC technique. Figure 3b orchestrates that an adversary is expected to make an additional 54m error when comparing our method with PPC. We observe that this error distance increases as ϵ decreases.

Entropy: Location entropy captures the uncertainty of the adversary during the inference of the correct location. Location entropy is given by:

$$H_l = -\sum P(x,y) \log_2 \left(P(x,y)\right) \tag{6}$$

where $P(x,y)$ is the probability that an object is located at position (x,y). We compared our method with the NWA technique for $\delta = 1000$. Since NWA does not anonymize the time domain, we left out the time domain of traces. We used the Geolife dataset to track a user for a given time and determine the uncertainty of the adversary for $\epsilon = 0.1, 0.5, 1.5$. Figure 3a depicts the results of the experiment. Our technique produced superior entropy results when compared to the NWA, despite the fact that our technique uses just a single object while NWA uses 20 and 5 moving objects. The Figure also shows that as ϵ reduces (meaning stronger privacy), the uncertainty or entropy increases. It is important to point out that our technique insert uncertainty to each trace of a trajectory and does not depend on neighboring objects (like in k-Anonymity). Thus, if traces of an outlier object is passed through our randomized mechanism with low ϵ values, a very strong privacy is guaranteed.

5.3 Quality and Utility of Perturbed Trace

Range Query Distortion: We evaluate the quality of our differential private perturbed traces using the range query distortion measure provided by NWA [1] which is given by:

$$\frac{\left|Q\left(\overline{\mathbf{HT}}\mathbf{r}_j\right) - Q\left(\mathbf{HTr}_j\right)\right|}{\max\left(Q\left(\overline{\mathbf{HT}}\mathbf{r}_j\right), Q\left(\mathbf{HTr}_j\right)\right)}$$

We used the same Oldenburg dataset and a similar query statement used in NWA [1] at Section 4C with $k = 140$. We used different privacy levels ϵ and a radius ranging from 300 to 4000. Figure 3c depicts the outcome. It shows that as ϵ increases, the range query distortion decreases. That is, as the privacy increases (low ϵ) more uncertainty is injected to prevent an adversary from inferring into the users privacy. In comparison to NWA, our approach shows a slightly lower distortion. In addition, only one moving object was used in this particular experiment to further emphasize that our approach protects outliers.

NNAR Evaluation: We conducted several experiments to evaluate the NNAR using the Athens dataset. There is no question about the benefits of context aware perturbed traces to database queries. However, to evaluate the benefits of context aware perturbed traces for data mining, we compared the utility of context aware traces produced by Algorithm 2 and the differential private high precision traces outputted by Algorithm 1 which are not context aware. We clustered each set of perturbed trace separately using KMeans and evaluated the quality of the cluster. Figure 3d and Figure 3e show the F1 measure results of the context aware trace and the differential private high precision trace, respectively. It can be seen that the F1 Measure for context aware NNAR released trace is better.

Runtime: The time required to perturb traces depends the type of the perturbed traces to be sent to the MOD or LBS, as well as the size of the Running Window. Context aware perturbation requires a longer time than its counterpart as depicted in Figure 3f. Figure 3f utilized the Athens dataset. This extra runtime overhead stems from the time needed to search for NNAR. Also, the higher the amount of traces per Running Window, the lower the runtime. This is because, the number of high precision traces generated will reduce if more raw GPS traces are partitioned into a Running window. Our proposed technique can be used for both stream data and non-stream datasets. For non-stream datasets, the server of our proposed technique requires minutes to perturb 90K traces.

6. CONCLUSIONS

We presented a novel technique to achieve differential privacy for stream and non-stream GPS data. Our technique utilizes a moving average filter to create high precision GPS traces and then perturbs these traces using a differential private randomized mechanism. We introduce the notion of Nearest Neighbor Anchor Resource, which ensures context aware location privacy by capturing and storing the location context of an object in an MOD or LBS, yet guaranteeing strong privacy. We orchestrate empirically that our technique protects outliers. Differential private RFID data protection is an interesting future work.

7. REFERENCES

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE*, 2008.

[2] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. di Vimercati, and P. Samarati. Location privacy protection through obfuscation-based techniques. In *DBSec'07*, pages 47–60, 2007.

[3] B. Belabbas, A. Hornbostel, M. Sadeque, and H. Denks. Accuracy study of a single frequency receiver using a combined gps/galileo constellation. In *ION GNSS*, 2005.

[4] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. KDD '10, 2010.

[5] A. Blum, C. Dwork, and K. Nissim. Practical privacy: The sulq framework. In *PODS'05*, 2005.

[6] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. STOC '08, pages 609–618, 2008.

[7] J. Bond. An investigation on the use of gps for deformation monitoring in open pit mines. *Geodesy and Geomatics Engineering*, 2004.

[8] G. Cormode, C. M. Procopiuc, D. Srivastava, E. Shen, and T. Yu. Differentially private spatial decompositions. In *ICDE*, pages 20–31, 2012.

[9] M. Duckham and L. Kulik. A formal model of obfuscation and negotiation for location privacy. In *Pervasive'05*, pages 152–170, 2005.

[10] C. Dwork. Differential privacy. In *ICALP*, 2006.

[11] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, pages 265–284, 2006.

[12] A. Friedman and A. Schuster. Data mining with differential privacy. KDD '10, pages 493–502, 2010.

[13] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. KDD '08, pages 265–273, 2008.

[14] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. ICDCS '05, pages 620–629, 2005.

[15] B. Hoh and M. Gruteser. Protecting location privacy through path confusion. In *SECURECOMM '05*, 2005.

[16] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias. Preventing location-based identity inference in anonymous spatial queries. *IEEE Trans. on Knowl. and Data Eng.*, 2007.

[17] J. Li, K. Miyashita, T. Kato, and S. Miyazaki. Gps time series modeling by autoregressive moving average method. *Earth Planets Space*, pages 155–162, 2000.

[18] S. Lim, T. Musa, and C. Rizos. Application of running average function to non-dispersive errors of network-based real-time kinematic positioning. In *Journal of Global Positioning Systems*, 2008.

[19] L. Liu. Privacy and location anonymization in location-based services. *SIGSPATIAL Special*, 2009.

[20] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE'08*, pages 277–286, 2008.

[21] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 2007.

[22] F. McSherry. Privacy integrated queries. In *SIGMOD '09*, 2009.

[23] F. Mcsherry and K. Talwar. Mechanism design via differential privacy. FOCS '07, 2007.

[24] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. Differentially private data release for data mining. KDD '11, pages 493–501, 2011.

[25] M. F. Mokbel. Query processing for location services without compromising privacy. In *VLDB'06*, 2006.

[26] M. E. Nergiz, M. Atzori, and Y. Saygin. Towards trajectory anonymization: a generalization-based approach. In *SPRINGL '08*, 2008.

[27] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC '07*, pages 75–84, 2007.

[28] D. Pfoser and C. S. Jensen. Capturing the uncertainty of moving-object representations. SSD '99, 1999.

[29] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. SIGMOD '10, 2010.

[30] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[31] M. Terrovitis and N. Mamoulis. Privacy preservation

in the publication of trajectories. In *MDM'08*, 2008.

[32] Y.-H. Tsai, F.-R. Chang, and W.-C. Yang. Moving average filters for faster gps receiver autonomous integrity monitoring. In *ION'02*, pages 666–675, 2002.

[33] V. S. Verykios and A. Gkoulalas. A free terrain model for trajectory k-anonymity. In *DEXA '08*, 2008.

[34] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *JASA*, pages 375–389, 2009.

[35] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu. Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In *ICDE*, 2008.

[36] Y. Zheng, Q. Li, Y. Chen, and X. Xie. Understanding mobility based on gps data. In *UbiComp*, 2008.

APPENDIX

A. SENSITIVITY BOUNDS

This section provides a proof of Lemma 1 in Section 3.4.

PROOF. Given that the moving average filter function $f(x) = \frac{1}{n}\sum_{k=1}^{n} x_k$ acts on a finite set A of i traces in the Running Window. Assume that $A = \{a_1, a_2, a_3, ...a_{i-1}, a_i\}$ where $|a_1| < |a_2| < |a_3| < ....|a_i|$. From this assumption, it implies the trace with the lowest magnitude $\min(A) = a_1$ while that with the highest magnitude $\max(A) = a_i$.

Given an average function $f(x)$, to get the (lower and upper) bounds of $f(x)$ in a Running Window, we remove the traces with the highest (a_i) and lowest (a_1) values . Removing a_1 or a_i from the set A results in the formation of two new sets. Namely, $A_{\overline{a1}} = \{a_2, a_3, ...a_{i-1}, a_i\}$, which corresponds to the set where the minimum trace value has been removed; and $A_{\overline{ai}} = \{a1, a2, a3, ...a_{i-1}\}$ is the set formed from the removal of the maximum trace value. Mathematically, when computing an average function $f(x)$ over a set of finite values, since $|a1| < |ai| \Rightarrow f(A_{\overline{a1}}) > f(A_{\overline{ai}})$.

This means that the maximum change will occur if the trace with the smallest magnitude is removed. Hence

$$\max(|f(\mathcal{T}_1) - f(\mathcal{T}_2)|)_{\forall_{A, A_{\overline{a1}}}} \leq \frac{1}{N-1}\sum_{m=1}^{N-1} A_m$$

$\forall m \neq \min(A)$

□