DP-WHERE: Differentially Private Modeling of Human Mobility

Darakhshan J. Mir*, Sibren Isaacman†, Ramón Cáceres‡, Margaret Martonosi§, Rebecca N. Wright*
*Rutgers University †Loyola University Maryland ‡AT&T Labs §Princeton University

***Abstract*—Models of human mobility have broad applicability in urban planning, ecology, epidemiology, and other fields. Starting with Call Detail Records (CDRs) from a cellular telephone network that have gone through a straightforward anonymization procedure, the prior WHERE modeling approach produces synthetic CDRs for a synthetic population. The accuracy of WHERE has been validated against billions of location samples for hundreds of thousands of cell phones in the New York and Los Angeles metropolitan areas. In this paper, we introduce DP-WHERE, which modifies WHERE by adding controlled noise to achieve *differential privacy*, a strict definition of privacy that makes no assumptions about the power or background knowledge of a potential adversary. We also present experiments showing that the accuracy of DP-WHERE remains close to that of WHERE and of real CDRs. With this work, we aim to enable the creation and possible release of synthetic models that capture the mobility patterns of real metropolitan populations while preserving privacy.**

I. Introduction

Models of human mobility have wide applicability to infrastructure and resource planning, analysis of infectious disease dynamics, ecology, and more. The abundance of spatiotemporal data from cellular telephone networks affords new opportunities to construct such models. Furthermore, such data can be gathered with greater detail at larger scale and lower cost than traditional methods, such as census surveys.

Prior work introduced the WHERE (Work and Home Extracted REgions) approach to mobility modeling [19]. In WHERE, aggregated collections of cellphone Call Detail Records (CDRs) form the basis of a mobility model that can be used to characterize a city's commute patterns and enable the exploration of what-if scenarios regarding changes in residential density, telecommuting popularity, etc. Starting with CDRs from a cellular telephone network that have gone through a straightforward anonymization procedure, WHERE produces synthetic CDRs for a synthetic population. WHERE has been experimentally validated against billions of location samples for hundreds of thousands of cell phones in the New York and Los Angeles metropolitan areas.

While human mobility models have the potential for great societal benefits, privacy concerns regarding their use of individuals' location data have inhibited their release and wider use. Although WHERE intuitively provides some privacy because it rests on aggregated distributions of sampled and anonymized data, a more rigorous assurance of privacy can further advance safe and widespread use of such techniques.

In this paper, we present and evaluate DP-WHERE, a *differentially private* version of WHERE. DP-WHERE satisfies the rigorous requirements of differential privacy while retaining WHERE's usefulness for predicting movement of human populations in metropolitan areas. Differential privacy [7] makes privacy a mathematical requirement on the results of interactions with data. In particular, differential privacy captures the intuition that, in order to provide privacy to individuals, the results of an interaction with a database should be almost the same whether or not any particular individual is present in a database. This is a strong notion of privacy that makes no assumptions about the power or background knowledge of a potential adversary.

DP-WHERE achieves differential privacy by adding controlled noise to the set of empirical probability distributions that WHERE uses, for example distributions of home and work locations. DP-WHERE then proceeds identically to WHERE by systematically sampling these distributions to generate synthetic CDRs containing synthetic locations and associated times. Because none of these sampling steps require further access to the original CDRs, it would be possible for the data holder to release the noisy distributions while retaining differential privacy. Among possible uses, these distributions would allow others to produce their own synthetic CDR traces for any desired population size, time duration, or other parameters.

Overall, our work shows that modest revisions to a mobility model drawn from real-world and large-scale location data allow for rigorous demonstrations of its privacy without overly compromising its utility. Specific contributions of our work include the following:

- We are, to our knowledge, the first to produce and evaluate a differentially private approach for modeling human mobility based on large sets of cellular network data.
- Our experiments show that differential privacy can be achieved with only a modest and acceptable reduction in accuracy. In particular, across a wide array of experiments involving 10,000 synthetic users moving across more than 14,000 square miles, the distance between synthetic and real population density distributions for DP-WHERE differed by only 0.17–2.2 miles from those of WHERE.
- More broadly, this work shows that there is reason for optimism regarding the judicious use of Big Data repositories of potentially sensitive information. We show the value of a multi-pronged approach to privacy: Our model starts with attributes (such as sampling and aggregation) that make it intrinsically well suited to offering some intuitive degree of privacy. We subsequently modify the steps of the modeling algorithm to rigorously implement differential privacy.

Fig. 1 outlines DP-WHERE and its changes to WHERE. We provide background on WHERE and differential privacy in Sec. II, describe DP-WHERE in Sec. III, evaluate its utility in Sec. IV, and discuss related work in Sec. V.

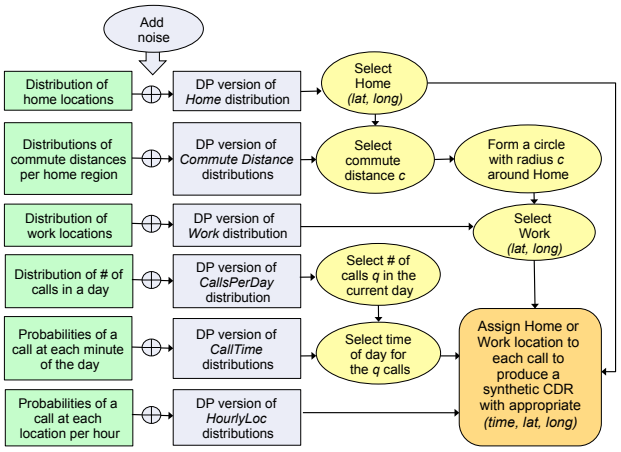

Fig. 1: Overview of DP-WHERE, which modifies WHERE by adding noise to achieve differentially private versions of the input probability distributions. The rest of WHERE remains unchanged.

II. Background

A. WHERE

DP-WHERE is based on WHERE, which produces models of how populations move within metropolitan areas [19]. WHERE generates sequences of times and locations that aim to capture how people move between important places in their lives, such as home and work. Previous work has shown that people spend most of their time at a few such places [12, 16, 30]. WHERE aggregates the movements of many synthetic individuals to reproduce human densities over time at the geographic scale of metropolitan areas.

WHERE draws information from either CDR traces or public sources (e.g., the US Census Bureau). It then creates a set of probability distributions that it uses to drive the generation of synthetic CDRs for the region being modeled. This paper uses as its starting point the version of WHERE that uses CDR traces as its data source. As shown in Fig. 6, this source yields substantially better experimental results than using current publicly available data sources.

The WHERE modeling algorithm takes as input a database of simplified CDRs. (Complete CDRs contain details not relevant to mobility, e.g., call-termination codes.) Each row of this database corresponds to a single voice call or text message, both of which we refer to interchangeably as calls. WHERE thus uses a database D of m entries corresponding to calls made by n distinct users. Each user is indexed by a unique anonymized user ID in the set $[n] = \{1, 2, \ldots n\}$. The calls were made in a given metropolitan area divided into smaller geographic areas by imposing a square grid of $d \times d$ cells.

WHERE leverages earlier work that estimates important places in people's lives (e.g., home and work) by applying clustering and regression methods to the CDRs in D [16]. In order to work with a single database in DP-WHERE, we append to each CDR entry these inferred home and work locations for the corresponding user. Thus, for the purposes of DP-WHERE, each row of D contains the following fields: `id`, `date`, `time`, `lat`, `long`, `home`, and `work`.

At its core, WHERE uses D to construct cumulative distribution functions (CDFs) for the following probability distributions (see also Fig. 1):

1) Home and Work: For each grid cell, all users with inferred `home` locations in that grid cell are counted (and normalized) to produce a probability distribution $Home$ over the grid cells. Similarly, a $Work$ distribution is constructed from the inferred `work` locations of users in the database.

2) CommuteDistance: WHERE allows for a coarser grid to be used for commute distances than for `home` and `work` locations by merging adjoining cells in the underlying $d \times d$ grid to yield a $d_c \times d_c$ grid. We refer to this coarser grid as the *commute grid*. For each cell in the commute grid, WHERE creates an empirical distribution of commute distances (i.e., distance between `home` and `work`) for people whose `home` locations are in that grid cell, leading to a total of d_c^2 of these $CommuteDistance$ distributions.

3) CallsPerDay: WHERE computes an empirical distribution $CallsPerDay$ over the set $\mathbb{C} = \{\mu_{\min}, \ldots \mu_{\max}\} \times \{\sigma_{\min}, \ldots \sigma_{\max}\}$ of possible rounded values of means and standard deviations of numbers of calls per day made by users.

4) ClassProb and CallTime: For each user in D, WHERE computes the distribution of when calls are made throughout the day. These per-user distributions are then combined using X-Means clustering into two classes [19]. Each user belongs to one of two user classes with a probability specified by $ClassProb$. Subsequently, using the CDR database, per-minute call probability distributions $CallTime$ are computed separately for each user class.

5) HourlyLocs: For each hour of the day, WHERE computes a distribution of calls made over the grid cells. Each of those 24 distributions reflects the probability of users being at a given location during that hour. The $HourlyLocs$ distributions are not tied to a specific user, but represent the calling activity across the entire metropolitan area during each hour.

As shown in Fig. 1, subsequent stages of WHERE use the above distributions to produce synthetic CDRs for any number of synthetic users and any length of simulated time. WHERE generates a synthetic user as follows. It first selects a `home` location by sampling from $Home$. It then selects a commute distance c by sampling $CommuteDistance$ for the commute-grid cell in which the home lies. Finally, it selects a `work` location by sampling from $Work$ while restricted to locations at distance c from the home location.

WHERE then generates synthetic call times and locations for a synthetic user i as follows. First, it samples from $CallsPerDay$ to obtain a (μ_i, σ_i) tuple that represents i's calling frequency. Second, it samples from the normal distribution with a mean μ_i and standard deviation σ_i to determine the number of calls q that i makes in the current simulated day. Third, it samples from $ClassProb$ to assign i one of two classes of calling time patterns. Fourth, it samples $CallTime$ to select the times of day for the q calls that day. Finally, it samples $HourlyLocs$ to determine the locations of these calls

while restricted to the user's home and work locations.

The synthetic CDR traces that comprise the output of WHERE have been shown to agree closely across a variety of metrics with real-world CDR traces for hundreds of thousands of users moving over metropolitan regions of thousands of square miles [19].

B. Differential Privacy

Differential privacy formalizes the idea that results of a data analysis should be almost the same whether or not an individual is in the database, even for individuals with unusual behaviors. This implies that the risk to an individual associated with a data analysis—including the risk of being identified—does not depend on whether the individual is in the database or not. This guarantee holds regardless of the external information available to an adversary. Differential privacy relies on the notion of neighboring databases [7]—in our context, two neighboring CDR databases. In our context, CDR databases D and D' are neighbors if they differ in the records of exactly one user (who may have made many calls).

Definition 1 (Neighbors)**.** *Two CDR databases D and D' are* neighbors *if $D \subset D'$ and there is some $k \in [n]$ such that for every record $r \in D' \oplus D$, $\texttt{id}(r) = k$ (where $\texttt{id}(r)$ denotes the user id in r).*

Definition 2 ([7])**.** *A randomized algorithm $\mathcal{A}$ is* ϵ-differentially private *if for all neighboring input data sets D, D' and for all $S \subseteq \mathrm{Range}(\mathcal{A})$, $\Pr[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D') \in S]$.*

That is, the output distributions for a differentially private $\mathcal{A}$ are similar for any two neighboring databases. The smaller the value of ϵ, the closer these two distributions are, and hence, the higher the privacy. The appropriate value for ϵ is a largely open policy question that depends on both privacy and accuracy needs. A wide range of values of ϵ, e.g., from 0.01 to 2.3, have been used in recent work [8, 22, 24].

An important notion in the application of differential privacy, the *global sensitivity* [7] of a function of a database is the maximum change in the value of the function over neighboring databases:

Definition 3 ([7])**.** *The* global sensitivity *of a function $f : D \to \mathbb{R}^\ell$ is $\mathrm{GS}_f := \max_{D,D'} \|f(D) - f(D')\|_1$, where D and D' are neighboring databases.*

One way to achieve differential privacy is to add noise to each element of the outcome of f that is proportional to the global sensitivity of f [7]. Specifically, let $\mathrm{Lap}(0, \lambda)$ denote a Laplace distribution with mean 0 and standard deviation $\sqrt{2}\lambda$, and let $\langle \mathrm{Lap}(0, \lambda) \rangle^\ell$ denote a length-ℓ vector of independent random samples from this distribution.

Theorem 1 ([7])**.** *For any $f : D \to \mathbb{R}^\ell$, and $\epsilon > 0$, the following mechanism $\mathcal{A}$, called the* Laplace mechanism*, is ϵ-differentially private: $\mathcal{A}_f(D) = f(D) + \langle \mathrm{Lap}(\mathrm{GS}_f / \epsilon) \rangle^\ell$.*

We make extensive use of the Laplace mechanism in DP-WHERE. Another generalized way of achieving differential privacy is the *exponential mechanism* [26]. Informally, the mechanism induces a probability distribution on $\mathrm{Range}(\mathcal{A})$ that is exponentially biased in favor of outputs that are closer to the real answer $f(D)$. Differential privacy for multi-step algorithms can be provided by breaking the algorithm down into multiple interactions with the database, each of which is itself differentially private. Thms. 2 and 3 formalize this:

Theorem 2 (Serial Composition [7])**.** *For $i \in [k]$, let $\mathcal{A}_i(D)$ be an ϵ_i-differentially private mechanism executed on database D. Then, any mechanism $\mathcal{A}$ that is a composition of $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$, is $\sum_i \epsilon_i$-differentially private.*

Theorem 3 (Parallel Composition [7])**.** *For $i \in [k]$, let $\mathcal{A}_i(D)$ be an ϵ_i-differentially private mechanism executed on partition D_i of the database D, such that $\forall i, j \ |D_i \cap D_j| = 0$, and each user appears in exactly one of the D_i's. Then, any mechanism $\mathcal{A}$ that is a composition of $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ is $\max_i \epsilon_i$-differentially private.*

III. Differentially Private WHERE

Our new approach, DP-WHERE, modifies WHERE to provide differential privacy while retaining the accuracy of the original WHERE approach. As described in Section II-A, WHERE creates and samples from several spatiotemporal distributions. In DP-WHERE, we render each of these distributions ϵ_i-differentially private, then apply Thms. 2 and 3 to arrive at an ϵ-differentially private algorithm, where $\epsilon = \sum_i \epsilon_i$. For each distribution, we specify a privacy budget ϵ_i that will not be exceeded. The remainder of this section describes DP-WHERE in detail.

A. Pre-processing

Before the algorithm executes, we perform a pre-processing step that removes all users who make more than a maximum threshold $\mathrm{MaxCallsHr}$ of calls per hour. This limits the impact of any one user on the dataset. Our experimental evaluation sets $\mathrm{MaxCallsHr}$ to 120 which makes it likely that any filtered caller is an auto-dialer; Section IV shows it yields good results.

B. Distributions

1) Home and Work: We compute differentially private empirical CDFs for $Home$ and $Work$. Let $\epsilon_{\texttt{home}}$ and $\epsilon_{\texttt{work}}$ be the privacy budgets allocated to computing $Home$ and $Work$, respectively. $\mathrm{CountHomeNum}(i)$ returns the number of distinct users in the database D with homes in the ith grid cell (in the chosen canonical ordering). Note that the global sensitivity (Def. 3) of $\langle \mathrm{CountHomeNum}(1), \ldots, \mathrm{CountHomeNum}(d^2) \rangle$ is 2, since each user can change his home location from grid cell i to another grid cell j, reducing the count in grid cell i by 1 and increasing j's count by 1. Applying the Laplace mechanism described in Thm. 1, Alg. III.1 provides an $\epsilon_{\texttt{home}}$-differentially private approximation of $Home$. Similarly, the Laplace mechanism achieves an $\epsilon_{\texttt{work}}$-differentially private CDF for $Work$.

The noisy "CDF" does not correspond to a legitimate probability distribution, as the noisy counts are not necessarily non-decreasing. We use Hay et al.'s post-processing techniques [13]

Algorithm III.1: $\textbf{DPHOMECDF}(D, \epsilon_{\texttt{home}})$

Count $\leftarrow 0$
for $i \leftarrow 1$ **to** d^2
do $\begin{cases} \text{Count} \leftarrow \text{Count} + \text{CountHomeNum}(i) + \\ \qquad \text{Lap}(0, \frac{2}{\epsilon_{\texttt{home}}}). \\ \text{CDF}[i] \leftarrow \text{Count}. \end{cases}$
CDF $\leftarrow$ PostProc(CDF)
output (CDF)

to clean up this noise and create a legitimate (non-decreasing) CDF, denoted by PostProc in Alg. III.1. The postprocessing method does not need to access the original private data, so Thms. 1 and 3 imply:

Lemma 1. *Alg. III.1 is ϵ_{home}-differentially private. The equivalent algorithm for Work is ϵ_{work}-differentially private.*

Fig. 2 shows the CDFs of the *Home* distribution for different values of $\epsilon_{\texttt{home}}$ and the original empirical CDF. (The dataset and parameters used for the figures are described in detail in Section IV.) The private version of *Home* is very close to its non-private counterparts even for very low values of $\epsilon_{\texttt{home}}$. Only for extreme values of ϵ such as 0.000001 are the differences even noticeable at this graph scale.

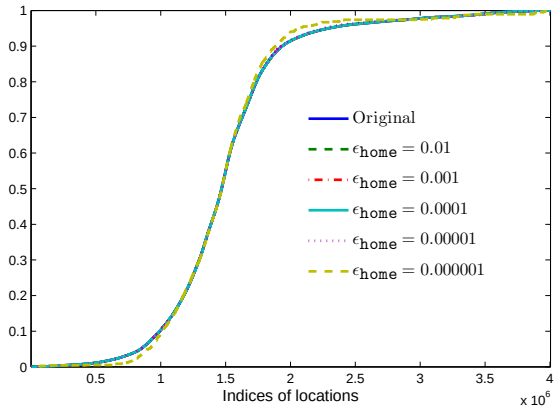

Fig. 2: CDF of *Home* distribution for different values of $\epsilon_{\texttt{home}}$.

2) Commute Distance: As in WHERE, we impose a $d_c \times d_c$ commute grid on the geographical area. We first create a data structure D_i that contains the counts of commute distances of users (in CDR database D) with `home` locations in grid cell i. We wish to avoid having empty grid cells, and to do so in a data-oblivious manner so that we do not spend more of our privacy budget. We therefore add two commute distances (of 0 and 0.1 miles) to every commute grid cell.

In WHERE, the CDF of a commute distribution is constructed by using the actual commute distances as histogram bins. In DP-WHERE, for privacy reasons, we cannot use the actual commute distances of a grid cell's residents as histogram bins. Instead, as shown in Alg. III.2, we create a per-commute-grid-cell data-dependent histogram of commute distances in a differentially private way, and then sample from this (normalized) histogram. The data-dependent histogram bins also need to be created in a differentially private manner. Let

Algorithm III.2: $\textbf{COMMUTECDFS}(D_i, i, \epsilon_{\texttt{commute}})$

CREATE BINS:
dpmedian $\leftarrow$ ExpoMedian$(D_i, \frac{\epsilon_{\texttt{commute}}}{2})$
synthdata $\leftarrow$ GenExpoSynthData(dpmedian)
bins $\leftarrow$ FindPercentiles(synthdata)

CREATE NOISY HISTOGRAM:
for $j \leftarrow 1$ **to** numbins
do $\begin{cases} \text{CDF}[i,j] \leftarrow \text{CountCommute}(\text{bins}_j, i) + \\ \text{Lap}(0, \frac{2 \cdot 2}{\epsilon_{\texttt{commute}}}). \end{cases}$
CDF$[i]$ $\leftarrow$ PostProc(CDF$[i]$)
output (CDF$[i]$)

$\epsilon_{\texttt{commute}}$ be the privacy budget for the commute distribution. We allocate half of this to determine the histogram bin ranges (because they are data dependent) and the other half to compute the counts themselves. To determine the bins, we assume that the commute distances in each grid cell are modeled by an exponential distribution—a popular model for positively skewed distributions such as commute distances, e.g., [1]. Let $\eta(x)$ be the (normalized) frequency of the distance x in the dataset D_i. If $\eta(x)$ follows an exponential distribution with rate parameter λ, then $\eta(x) = \lambda e^{-\lambda x}$. The rate parameter can be estimated using the median of the empirical data, by $\hat{\lambda} = \text{median} / \log(2)$. The differentially private approximation to the median of the commute distances in grid cell i is called dpmedian and is computed using a computationally efficient version of the *exponential mechanism* [26], as in [5]. In Alg. III.2, ExpoMedian$(D_i, \frac{\epsilon_{\texttt{commute}}}{2})$ implements this algorithm to compute dpmedian, an $\frac{\epsilon_{\texttt{commute}}}{2}$-differentially private approximation of the median of the commute distances.

Next, we determine the histogram bins by creating a large synthetic set of commute distances that are sampled from an exponential distribution whose parameter is given by $\lambda = \frac{\text{dpmedian}}{\log(2)}$. In Alg. III.2, GenExpoSynthData(dpmedian) generates a set of synthetic commute distances, synthdata, from such a distribution. We determine the $10, 20, 30, \ldots, 90, 95$ percentiles of this set of distances using FindPercentiles. The distances corresponding to these percentiles form the edges of the histogram bins.

CountCommute(bins_j, i) counts the number of distances in the data structure D_i that fall in bins_j. $\langle \text{CountCommute}(\text{bins}_1, j), \ldots, \text{CountCommute}(\text{bins}_{10}, i) \rangle$ has a global sensitivity of 2. Applying the Laplace mechanism yields an $\frac{\epsilon_{\texttt{commute}}}{2}$-differentially private computation of the approximate histogram counts. Since each user appears in only one of the $d_c \times d_c$ grid cells, by Thms. 2 and 3 and the privacy of the ExpoMedian [5]:

Lemma 2. *Using Alg. III.2 to compute* commuteCDF$(D_i, i, \epsilon_{commute})$, $\forall i \in \{1, \ldots, d_c^2\}$ *is $\epsilon_{commute}$-differentially private.*

3) Calls per Day per User: To create the CDF of *CallsPerDay* in a differentially private manner, we begin, as in WHERE, by assuming that the average number of calls per day for any user is from the set $\mathbb{M} = \{\mu_{\min}, \ldots, \mu_{\max}\}$.

Similarly, the standard deviation of the number of calls per day is from the set $\Sigma = \{\sigma_{\min}, \ldots, \sigma_{\max}\}$. Just as for WHERE, each μ_i and σ_i corresponding to a user i is rounded to the nearest value in the sets $\mathbb{M}$ and Σ, respectively.

Algorithm III.3: CALLSPERDAYCDF$(D, \epsilon_{\texttt{cpday}})$

COUNT:
for $\mu \leftarrow \mu_{\min}$ **to** μmax
do { **for** $\sigma \leftarrow \sigma_{\min}$ **to** σmax
do $\{\widehat{\mathrm{M}}(\mu, \sigma) \leftarrow \mathrm{CountAvgStd}(\mu, \sigma)$.

NOISE ADDITION:
for $\mu \leftarrow \mu_{\min}$ **to** μmax
do { **for** $\sigma \leftarrow \sigma_{\min}$ **to** σmax
do { $\widehat{\mathrm{M}}(\mu, \sigma) \leftarrow \widehat{\mathrm{M}}(\mu, \sigma) +$ Lap$(0, \frac{2}{\epsilon_{\texttt{cpday}}})$.

CONVERT TO CDF:
CDF $\leftarrow$ PostProc$(\widehat{\mathrm{M}})$
output (CDF)

Let $\mathrm{CountAvgStd}(\mu, \sigma)$ be a function that counts the number of users whose calls made per day have a (rounded) mean and standard deviation of μ and σ respectively. Consider the matrix M, of size $|\mathbb{M}| \times |\Sigma|$, whose elements corresponds to $\mathrm{CountAvgStd}(\mu, \sigma)$, for $\mu \in \mathbb{M}$ and $\sigma \in \Sigma$. Any addition or deletion of calls by a single user can change the mean / standard deviation pair from (μ, σ) to another pair (μ', σ'), decreasing the count for at most one element of the matrix M by at most 1 and increasing the count for another element by 1. Therefore, the global sensitivity of the vector $\langle M(\mu_{\min}, \sigma_{\min}), \ldots M(\mu_{\max}, \mu_{\min})\rangle$ is 2.

Alg. III.3 first counts each user's (μ, σ). At the end of the **COUNT** process in Alg. III.3, element $\widehat{\mathrm{M}}(\mu, \sigma)$ contains $\mathrm{CountAvgStd}(\mu, \sigma)$, $\forall \mu \in \mathbb{M}$ $\sigma \in \Sigma$. Using Thms. 1 and 3, the computation of $\widehat{\mathrm{M}}$ after it goes through **NOISE ADDITION** is differentially private. Next, the noisy matrix $\widehat{\mathrm{M}}$ is converted to a CDF by applying post-processing techniques [13] to further reduce the noise. Fig. 3 shows the differentially private approximation of the CDF of the *CallsPerDay* distribution for different values of $\epsilon_{\texttt{cpday}}$.

Lemma 3. *Alg. III.3's computation of* $\widehat{\mathrm{M}}$ *and the CDF of the CallsPerDay distribution is* $\epsilon_{\texttt{cpday}}$*-differentially private.*

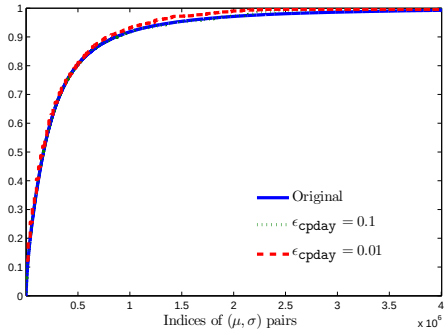

Fig. 3: CDF of *CallsPerDay*

Algorithm III.4: DP-KMEANS$(P, \epsilon_{\texttt{bdg}}, \epsilon_{\texttt{it}}, \texttt{tol})$

INITIALIZE:
$\mathrm{ClustCtr}_1 \leftarrow \langle \mathrm{Rand}\rangle^{24}$
$\mathrm{ClustCtr}_2 \leftarrow \langle \mathrm{Rand}\rangle^{24}$
$\epsilon_{\texttt{calltime}} \leftarrow 0$
ITERATE:
while $\epsilon_{\texttt{calltime}} \leq \epsilon_{\texttt{bdg}}$ **or** $\texttt{err} < \texttt{tol}$
do {
OldCtr $\leftarrow$ ClustCtr$_1$
OldCtr $\leftarrow$ ClustCtr$_2$
ClustSize$_1 \leftarrow$ ClustSize$_1 +$ Lap$(0, \frac{1}{\epsilon})$
ClustSize$_2 \leftarrow$ ClustSize$_2 +$ Lap$(0, \frac{1}{\epsilon})$
$\epsilon_{\texttt{calltime}} \leftarrow \epsilon_{\texttt{calltime}} + \epsilon_{\texttt{it}}$
Sum$_1 \leftarrow$ Sum(Cluster$_1$) $+ \langle$Lap$(0, \frac{2}{\epsilon})\rangle^{24}$
Sum$_2 \leftarrow$ Sum(Cluster$_2$) $+ \langle$Lap$(0, \frac{2}{\epsilon})\rangle^{24}$
$\epsilon_{\texttt{calltime}} \leftarrow \epsilon_{\texttt{calltime}} + \epsilon_{\texttt{it}}$
ClustCtr$_1 \leftarrow$ Sum$_1$ / ClustSize$_1$
ClustCtr$_2 \leftarrow$ Sum$_2$ / ClustSize$_2$
ClustCtr$_1 \leftarrow$ PostProc(ClustCtr$_1$)
ClustCtr$_2 \leftarrow$ PostProc(ClustCtr$_2$)
$\texttt{err} = \mathrm{dist}(\mathrm{OldCtr}_1, \mathrm{ClustCtr}_1)$
$+ \mathrm{dist}(\mathrm{OldCtr}_2, \mathrm{ClustCtr}_2)$

output (ClustSize$_1$, ClustSize$_2$)
output (ClustCtr$_1$, ClustCtr$_2$, $\epsilon_{\texttt{calltime}}$)

4) Call Times per User Class: In DP-WHERE, we cluster users into one of the two classes using differentially private k-means clustering [25] (rather than X-means as used in WHERE). From the CDR database D, just as in WHERE, we compute the number of calls each user makes during each hour of the day. From this, a 24-dimension probability vector (one dimension for each hour) is constructed so that each element represents the probability that a user makes calls during that hour. We classify users based on this 24-dimension probability vector. An intermediate data structure P that is input for the clustering algorithm (Alg. III.4) is the set of probability vectors p_i for all users i. Each row of P consists of the id of the user and his calling probability vector p_i. The input to Alg. III.4 consists of P, the target number k (2 in our work) of cluster centers, the privacy budget for the clustering algorithm $\epsilon_{\texttt{bdg}}$, the amount of the privacy budget $\epsilon_{\texttt{it}}$ that is spent for each iteration within the clustering algorithm, and the error tolerance $\texttt{tol}$. Alg. III.4 will iterate until either the error is within the range of tolerance or the privacy budget is used up.

As shown in Alg. III.4, we initialize the cluster centers by picking two random 24-dimension probability vectors $\langle \mathrm{Rand}\rangle^{24}$. Vectors in P are assigned to a cluster depending on which of the two current cluster centers they are closer to. Over each iteration, the noisy sum of the vectors in $\mathrm{ClustCtr}_i$ for each current cluster i is computed. The global sensitivity of the sum of vectors in $\mathrm{ClustSum}_i = \sum_{j \in \mathrm{Cluster}_i} p_j$ is 2, because $\forall j \in [n], \|p_j\|_1 = 1$ (since each of the vectors is a probability vector). Any change in one person's data can change $\mathrm{ClustSum}_i$ to another vector $\mathrm{ClustSum}_i + \delta$, where $|\delta| \leq 2$. The size of a cluster has global sensitivity 1. A

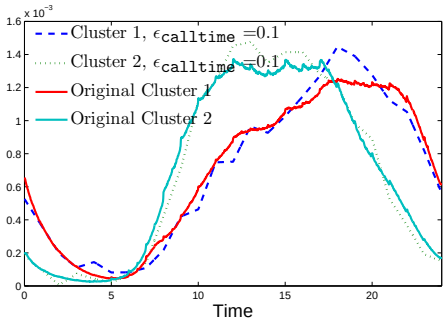

Fig. 4: Comparison of distribution of call times for two classes of users as determined by Alg. III.4 to the non-private clustering.

differentially private computation of the cluster size and sum enables a differentially private approximation of the mean vector of the cluster. For each iteration, the computation of each ClustSize_i, as well of Sum_i, is $\epsilon_{\texttt{it}}$-differentially private. By Thm. 2, this leads to a $2\epsilon_{\texttt{it}}$-differentially private computation of ClustCtr_i. By Thm. 3, the privacy level for an iteration is also $2\epsilon_{\texttt{it}}$, as the clusters are non-intersecting subsets of the dataset P. A user—and consequently his probability vector—appears in exactly one cluster.

At this point, ClustCtr_i, the noisy mean of the vectors in Cluster i, will not necessarily correspond to a probability vector, as some of its elements may be negative and their sum may not add to 1. To correct for this, we apply post-processing noise correction techniques on each of these cluster centers before returning to the next iteration. After an iteration where either the privacy budget is exhausted or the error falls below the given threshold `tol`, the algorithm returns the differentially private cluster centers, the total privacy budget spent ($\epsilon_{\texttt{calltime}}$), and a differentially private computation of the cluster sizes (the vector ClustSize). All of this incurs a privacy expenditure of $\epsilon_{\texttt{calltime}}$.

We use the cluster centers as calling time probability distributions: each element of the cluster center vectors represents the probability that a user in that cluster makes a call during that hour. We compute one probability distribution $CallTime$ for each minute of the day and for each user class by interpolating the probability distribution over all minutes between the hours (elements of the cluster centers). We use ClustSize to determine $ClassProb$, the probability of a user belonging to one of the two classes. Using Thms. 2 and 3:

Lemma 4. *Alg. III.4 gives an $\epsilon_{\texttt{calltime}}$-differentially private clustering of the user calling probability vectors.*

Fig. 4 shows that DP-WHERE preserves typical diurnal patterns for both classes, even for low values of $\epsilon_{\texttt{calltime}}$.

5) Hourly Calls per Location: For every hour of the day, DP-WHERE differentially privately computes a empirical distribution of calls made over every grid cell. To do this, $\text{CountCallsNum}(i, j)$ is defined as the function that returns the number of calls users in D make in the ith grid cell between the hour $j-1$ and j. We wish to determine a matrix H of size $d^2 \times 24$; each row of H corresponds to a grid cell $i \in [d^2]$ and each column to an hour $j \in [24]$. Element $H(i, j)$ of the matrix has value $\text{CountCallsNum}(i, j)$. Let NumDays be the number of days that the database D corresponds to. The (column) vector corresponding to calls made over the geographical area during hour j is written as $\langle H(*, j)\rangle = \langle H(1, j) \ldots H(d^2, j)\rangle$. Since any change in exactly one user's data can cause a change of at most MaxCallsHr for every hour of each of these days, the global sensitivity of this vector is $\text{MaxCallsHr} \cdot \text{NumDays}$.

Algorithm III.5: HOURLYCDFS$(D, \epsilon_{\texttt{hrlocs}})$

$\text{gnums} \leftarrow \lfloor \frac{d^2}{\text{gsize}} \rfloor$
for $j \leftarrow 1$ **to** 24
  do
  for $\ell \leftarrow 1$ **to** gnums
    do
  GROUP:
  $g_\ell \leftarrow 0$
  for $i \leftarrow 1$ **to** gsize
    do
  $g_\ell \leftarrow g_\ell + \text{CountCallsNum}(i, j)$
  NOISE ADDITION:
  $\langle g\rangle \leftarrow \langle g\rangle + \left\langle \text{Lap}(0, \frac{\text{MaxCallsHr} \cdot \text{NumDays}}{\epsilon_{\texttt{hrlocs}}/24}) \right\rangle^{\text{gnums}}$
  RECONSTRUCT:
  $\langle H(*, j)\rangle \leftarrow \text{Reconstruct}(\langle g\rangle)$
  $\text{CDF}[j] \leftarrow \text{PostProc}(\langle H(*, j)\rangle)$
output (CDFs)

Direct use of the Laplace mechanism with this level of global sensitivity would add a lot of noise relative to the individual counts. To reduce the overall magnitude of noise added, we make use of *grouping* [20], which groups similar counts together and allows the magnitude of the noise added to each group count to be lower as compared to the total group count. Specifically, we set the group size gsize to be equal to $24 \cdot \text{NumDays}$. This is comparable to the magnitude of noise we will add to the resulting grouped-counts vector. Grouping gsize contiguous elements together yields a vector $\langle g\rangle$ of size $\text{gnums} = \lfloor \frac{d^2}{\text{gsize}} \rfloor$. Each element g_ℓ of $\langle g\rangle$ counts the total number of calls made in locations that appear in group ℓ. Note that the global sensitivity of $\langle g\rangle$ is still $\text{MaxCallsHr} \cdot \text{NumDays}$ because any one user can make up to a maximum of MaxCallsHr calls during a particular hour of each of these days. We then apply the Laplace mechanism to add noise to each group count. Finally, we replace every individual count $H(i, j)$ by the average of the noisy group count it belongs to (as denoted by Reconstruct in Alg. III.5).

Alg. III.5 applies a similar grouping scheme for each hour $(1, ..., 24)$. By Thm. 1, each of these computations is $\frac{\epsilon_{\texttt{hrlocs}}}{24}$-differentially private. Thus by Thm. 2:

Lemma 5. *Alg. III.5 is $\epsilon_{\texttt{hrlocs}}$-differentially private.*

While this kind of grouping may not always yield highly accurate results, in our case each of the hourly distributions is defined on a geographical area, so we can expect call counts within a group (corresponding to call counts in contiguous ge-

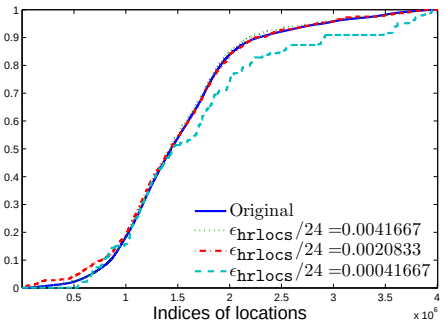

Fig. 5: *HourlyLocs* Distribution for 5:00pm to 6:00pm

ographical areas) to be similar to each other for many groups. As demonstrated by Fig. 5, showing *HourlyLocs* for different values of $\epsilon_{\texttt{hrlocs}}/24$, corresponding to an overall $\epsilon_{\texttt{hrlocs}}$ (over all the *HourlyLocs* distributions) of 0.1, 0.05, and 0.01, respectively, this method works well in our experiments.

C. DP-WHERE: Putting It All Together

The approximations to the empirical distributions computed above are ϵ_i-differentially private for different values of ϵ_i. DP-WHERE composes these individual differentially private mechanisms to yield the overall algorithm. To generate synthetic CDRs from these distributions, DP-WHERE performs the same steps as WHERE to sample from each of these private distributions to generate synthetic CDRs without going back to the original data. Applying Thm. 2 to Lem. 1–5 yields:

Theorem 4. *DP-WHERE is ϵ-differentially private, where*

$$\epsilon = \epsilon_{home} + \epsilon_{work} + \epsilon_{commute} + \epsilon_{cpday} + \epsilon_{calltime} + \epsilon_{hrlocs}.$$

It is important to note that, because none of the sampling steps in DP-WHERE require further access to the original data, it is possible for the data holder to release the noisy distributions while retaining differential privacy. This would allow others to produce their own synthetic CDR traces for any desired number of users, time duration, or other parameters.

IV. Experimental Evaluation

We have shown that DP-WHERE achieves differential privacy. Because it achieves this by injecting noise, we must also assess the impact on utility. In this section, we explore this impact by comparing the utility of the models produced by DP-WHERE and by WHERE. Specifically, for multiple uses, we demonstrate that DP-WHERE achieves similar accuracy to WHERE using real CDRs as input, and far better accuracy than WHERE using public data (e.g., the US Census) as input.

A. Datasets and Methodology

The input data for our DP-WHERE and WHERE experiments come from a large set of CDRs generated by actual cellphone use over 91 consecutive days from April 1 to June 30, 2011. This dataset contains over 1 billion records for both voice calls and text messages involving over 250,000 unique phones chosen at random from phones billed to ZIP codes within 50 miles of the center of New York City.

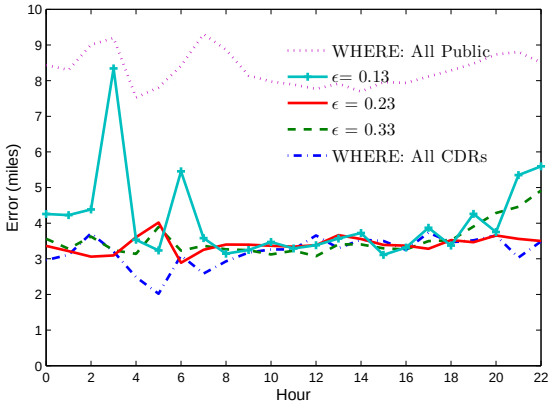

Fig. 6: EMD error for DP-WHERE using different values of ϵ and a fixed commute-grid cell size of 0.01° × 0.01°, as compared to WHERE using CDRs and WHERE using public data.

In addition to the differential privacy provided by DP-WHERE, we took several steps to preserve the privacy of individuals represented in our input datasets throughout our handling of those datasets. First, we used only anonymized CDRs containing no Personally Identifying Information (PII). Second, we did not focus our analysis on any individual phone. Third, we present only aggregate results.

In each of our DP-WHERE and WHERE experiments, we generate 10,000 synthetic users that travel for 30 consecutive days in an area of more than 14,000 mi^2 around New York City, more specifically bounded by latitudes 40°N & 42°N and longitudes 73°W & 75°W. This area is further broken down into squares 0.001° on a side to construct the $d \times d$ grid discussed in Section II-A, with d = 2,000.

B. Earth Mover's Distance

An important goal of our modeling approach is that a synthetic CDR trace should produce population density distributions that closely match those produced by a real CDR trace at every time of day. We therefore need a quantitative measure for comparing two spatial probability distributions. Our chosen metric is Earth Mover's Distance (EMD) [29].

EMD finds the minimum amount of energy required to transform one probability distribution into another. If one visualizes the problem as reshaping one mound of earth to match another, this energy is given by the amount of probability to be moved and the distance to move it. Thus, a lower EMD value indicates a stronger similarity between two distributions. Since different distance weightings lead to different EMD values, we follow the method in [19] and convert a raw EMD value to miles of error by using a normalizing factor. We obtain this factor by calculating the EMD between two spatial probability distributions with their entire populations concentrated in one of two places one mile apart.

The differential privacy parameter ϵ gives us a knob by which to trade privacy for accuracy. Figure 6 compares DP-WHERE using different values of ϵ to WHERE using CDRs and WHERE using public data. The size of the commute-grid cells is held constant at 0.01° × 0.01°. As shown, WHERE using CDRs has the lowest overall EMD, but DP-

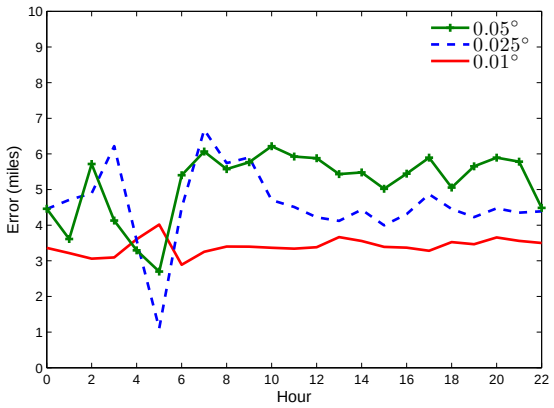

Fig. 7: EMD error for DP-WHERE using different sizes of commute-grid cell size and a fixed ϵ of 0.23.

WHERE performs favorably across a range of ϵ values, always performing better than WHERE using public data. As ϵ is made smaller to achieve better privacy, more noise is added and the EMD creeps upward. Accuracy is better in some times of day than in others. In particular, fewer people make calls in hours before 8 or after 22, so there is a smaller sample of locations in the input CDRs and adding noise has more of an impact during those hours.

The accuracy of DP-WHERE also depends on the granularities of the grids used to divide the geographic region of interest. Figure 7 compares DP-WHERE using different commute-grid sizes for the same ϵ of 0.23. At this ϵ value, coarser commute grids provide less accurate EMD results.

	commute-grid cell size		
	0.01°	**0.025°**	**0.05°**
WHERE	3.2150	3.3396	3.0871
ϵ **= 0.33**	3.5316	3.1655	4.5687
ϵ **= 0.23**	3.4066	4.5577	5.1691
ϵ **= 0.13**	5.3391	5.3194	5.2754

TABLE I: Average EMD error for WHERE using CDRs and DP-WHERE using various ϵ, as the commute-grid cell size changes.

We ran a wide range of experiments to explore the effects of ϵ and commute-grid cell size. Table I summarizes the EMD error averaged over the 24 hours of the day for each choice of ϵ and cell size. Across all our experiments, the EMD error for DP-WHERE differ by only 0.17–2.2 miles from those of WHERE using CDRs as input. Although EMD errors of 3 miles may appear large, note that EMD is aggregated over the entire area (i.e., over more than 14,000 square miles) and WHERE has already been validated to work with similar EMD [19]. As we further demonstrate in the following section with daily range, differential privacy can be achieved for a modest reduction in accuracy that still allows for useful mobility studies.

C. Daily Range

Daily range, or the maximum distance between any two points a person visits in a day, has proven useful for characterizing human mobility patterns [17, 18, 21]. We can therefore demonstrate the value of our modeling techniques by showing that the daily range computed from DP-WHERE's synthetic CDRs closely match those computed from real CDRs.

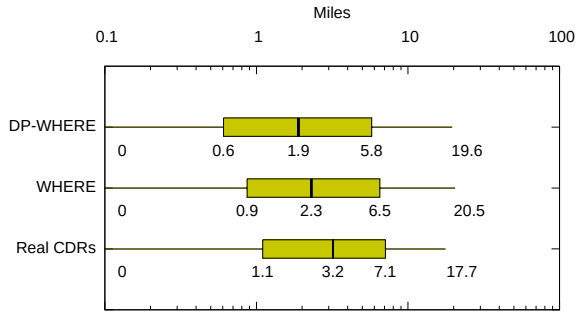

Fig. 8: Daily range for DP-WHERE ($\epsilon = 0.23$, commute-grid size $= 0.01° \times 0.01°$), WHERE from CDRs, and the real CDR dataset.

Figure 8 demonstrates the utility of DP-WHERE for daily range experiments. We compare daily ranges produced by DP-WHERE, WHERE from CDRs, and the original CDRs. We use boxplots to summarize the resulting empirical distributions, where the box represents the 25th, 50th, and 75th percentiles, while the whiskers indicate the 2nd and 98th percentiles. The horizontal axis shows miles on a logarithmic scale. Like WHERE, DP-WHERE exhibits daily ranges that are qualitatively similar to those from real CDRs, with differences of 0.5–1.3 miles across the middle two quartiles.

EMD and daily range serve as important and complementary metrics for validating our synthetic models. EMD measures the aggregate behavior of synthetic users, while daily range yields results at a per-user granularity. In summary, our EMD and daily range results confirm that DP-WHERE produces synthetic CDRs that closely mimic the behavior of large populations of real cellphone users.

V. Related Work

Mobility Modeling: Characterizing human mobility based on cellular network or other position data has received considerable attention. Our prior work developed algorithms for estimating people's daily range of travel [17, 18] and for inferring important locations in people's lives [16] from anonymized cellular network data. We further used this information to characterize commute distances, to quantify carbon footprints, and to create WHERE models [16, 19].

Early mobility modeling work used either handheld GPSs or WiFi associations to model human mobility at much finer scales, and with little privacy [15, 21, 28]. Prior uses of cellular data also include some mobility modeling [9, 10, 12, 30], but with little attention to privacy assurances. Such studies use at most anonymization and aggregation, and in some cases point to data characteristics that increase the difficulty of creating privacy-preserving mobility models.

Privacy: To our knowledge, the problem of creating differentially private human mobility models based on real-world cellular network data has not been studied previously. Differential privacy has been examined in other contexts of spatio-temporal data. Chen et al. [3] study the problem of publishing a differentially private version of the trajectory data of commuters in Montreal. They then evaluate the utility of published private data in terms of count queries and frequent

sequential pattern mining. Similarly, [6] recently characterized sequences of movements from individual users and found them extremely resistant to privacy techniques. In contrast, WHERE does not directly model the sequentiality of the spatio-temporal data at the level of an individual. Some work [14, 27] considers aspects of differential privacy on spatial data, but without DP-WHERE's end-to-end treatment. Other characterization work also exists. Several recent papers have characterized the privacy risks of releasing location data, in each case demonstrating the ability to re-identify individual information from geospatial data sets [11, 23, 31]. These papers motivated us to look beyond a simple anonymization of location traces. In addition, Andrés et al. [2] introduce the notion of *geo-indistinguishability* in location-based systems, which protects the exact location of a user while allowing release of information needed to gain access to a service.

VI. Conclusions and Future Work

DP-WHERE provides differential privacy while maintaining the utility of WHERE for modeling human mobility from real-world cellular network data. Our work shows it is possible to balance privacy and utility in practical big data applications.

Extensions to DP-WHERE could further improve its already strong accuracy. In particular, we note that the phones used in this study were sampled from a much larger set of millions of phones in each metropolitan area of interest. In our case, independent random samples were drawn from phones billed to each ZIP code in a metropolitan area. If we instead sample uniformly over all phones in a metropolitan area, we can improve the overall privacy guarantees of our algorithm, as noted by Cormode et. al [4]. For example, sampling 5% of phones from a database of millions of phones and running our DP-WHERE algorithm over the calls these phones generate would yield an order of magnitude improvement in the privacy parameter ϵ. Conversely, this sampling could be used to achieve a given ϵ with much less noise addition.

We hope that DP-WHERE constitutes a significant step towards enabling cellular telephony providers to unlock the value of their location data for applications with broad societal benefits, such as urban planning and epidemiology, without compromising privacy.

Acknowledgements. This work was supported in part by the National Science Foundation under grant numbers CCF-1018445 and CNS-1135953. Martonosi's work is funded in part by the Intel Science and Technology Center for Cloud Computing (ISTC-CC). We thank Graham Cormode for valuable comments and suggestions.

References

[1] J. Anable, C. Brand, M. Tran, and N. Eyre, "Modelling transport energy demand: A socio-technical approach," *Energy Policy*, vol. 41, pp. 125 – 138, 2012, Modeling Transport (Energy) Demand and Policies.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," *CoRR '12*, vol. abs/1212.1984.

[3] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: a case study on the Montreal transportation system," in *KDD '12*.

[4] G. Cormode, C. M. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *ICDE '12*.

[5] G. Cormode, C. M. Procopiuc, D. Srivastava, and T. T. L. Tran, "Differentially private summaries for sparse data," in *ICDT '12*.

[6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, Mar 2013.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC '06*.

[8] ——, "Differential privacy—a primer for the perplexed," in *Joint UNECE/Eurostat work session on statistical data confidentiality '11*.

[9] F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, and J. Blat, "Uncovering the presence and movements of tourists from user-generated content," in *Intn'l Forum on Tourism Statistics*, 2008.

[10] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate," in *CUPUM '09*.

[11] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Pervasive '09*.

[12] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, 2008.

[13] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," *PVLDB '10*, vol. 3, no. 1, pp. 1021–1032, 2010.

[14] S.-S. Ho and S. Ruan, "Differential privacy for location pattern mining," in *SPRINGL '11*.

[15] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks," in *INFOCOM '07*.

[16] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive '11*.

[17] ——, "Ranges of human mobility in Los Angeles and New York," in *MUCS '11*.

[18] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky, "A tale of two cities," in *HotMobile '10*.

[19] S. Isaacman, R. A. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *MobiSys '12*.

[20] G. Kellaris and S. Papadopoulos, "Practical differential privacy via grouping and smoothing," in *VLDB '13*.

[21] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *INFOCOM '06*.

[22] A. Korolova, "Protecting privacy while mining and sharing user data," Ph.D. Thesis, 2011.

[23] J. Krumm, "Inference attacks on location tracks," in *Pervasive '07*.

[24] Y. Lindell and E. Omri, "A practical application of differential privacy to personalized online advertising." *IACR Cryptology ePrint Archive*, 2011.

[25] F. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," *CACM*, vol. 53, no. 9, pp. 89–97, 2010.

[26] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS '07*.

[27] W. H. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," in *ICDE, '13*.

[28] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility: Do humans walk like monkeys?" in *INFOCOM '08*.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, 2000.

[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, 2010.

[31] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *MobiCom '11*.