Human Mobility Modeling at Metropolitan Scales

Sibren Isaacman*, Richard Becker†, Ramón Cáceres†, Margaret Martonosi*,
James Rowland†, Alexander Varshavsky†, Walter Willinger†
*Princeton University, Princeton, NJ, USA †AT&T Labs, Florham Park, NJ, USA
isaacman@princeton.edu, {rab,ramon}@research.att.com, mrm@princeton.edu,
{jrr,varshavsky,walter}@research.att.com

ABSTRACT

Models of human mobility have broad applicability in fields such as mobile computing, urban planning, and ecology. This paper proposes and evaluates *WHERE*, a novel approach to modeling how large populations move within different metropolitan areas. WHERE takes as input spatial and temporal probability distributions drawn from empirical data, such as Call Detail Records (CDRs) from a cellular telephone network, and produces synthetic CDRs for a synthetic population. We have validated WHERE against billions of anonymous location samples for hundreds of thousands of phones in the New York and Los Angeles metropolitan areas. We found that WHERE offers significantly higher fidelity than other modeling approaches. For example, daily range of travel statistics fall within one mile of their true values, an improvement of more than 14 times over a Weighted Random Waypoint model. Our modeling techniques and synthetic CDRs can be applied to a wide range of problems while avoiding many of the privacy concerns surrounding real CDRs.

Categories and Subject Descriptors

I.6.5 [**Computing Methodologies**]: Simulation and Modeling—*Model Development*; K.4.0 [**Computing Milieux**]: Computers and Society—*General*

General Terms

Algorithms, Measurement, Human Factors

Keywords

Human mobility patterns, Call Detail Records

1. Introduction

Human mobility models have myriad uses in mobile computing research and other fields of study. Models that faithfully reproduce the movements of real people can help answer questions in areas as varied as mobile sensing, opportunistic networking, urban planning, ecology, and epidemiology. For example, a model of how people move around a city can help evaluate whether a sensing application running on mobile phones would be able to attain the desired geographic coverage.

Our work aims to produce accurate models of how large populations move within different metropolitan areas. In pursuit of this general aim, we have a number of more specific goals. Our first goal is to generate sequences of locations and associated times that capture how individuals 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 [13, 16, 31]. Our second goal is to aggregate the movements of many such individuals to reproduce human densities over time at the geographic scale of metropolitan areas. A model that operates at these scales can help address important societal issues such as the environmental impact of home-to-work commutes. Our third goal is to take into account how different metropolitan areas exhibit distinct mobility patterns due to differences in geographic distributions of homes and jobs, transportation infrastructures, and other factors. Previous work has shown significant differences between cities along metrics such as commute distances [16, 17, 18, 25].

Many human mobility models that fall short on one or more of these goals have been proposed in the past. Some models produce random motion that does not correspond to actual mobility patterns, e.g., [19, 24]. Their lack of memory about recurring movement patterns and of spatiotemporal realism about population densities results in unrealistic motion of modeled individuals. Some models are tailored to a small geographic area such as a university campus, e.g., [20]. They do not apply to larger geographic areas with more diverse populations. Some models aim to be universal, e.g., [13], and thus do not adapt to different geographic areas. There remains a need for a realistic model that matches empirical observations for large and distinct geographic areas.

This paper introduces a modeling approach that takes as input certain spatial and temporal probability distributions drawn from large populations of real people living across wide geographic areas. An especially good source of these distributions are the Call Detail Records (CDRs) maintained by cellular network operators. Billions of cellphone users worldwide keep their phones near them most of the time, and the networks need to know the rough location of all active phones to provide them with voice and data services. CDRs contain information such as the time a voice call was placed or a text message was received, as well as the identity of the cell tower with which the phone was associated at that time. When joined with information about the locations of those towers, CDRs can serve as sporadic samples of the approximate locations of the phone's owner. A growing body of work has shown that information derived from anonymized CDRs can accurately characterize many aspects of human mobility [3, 11, 13, 16, 17, 18, 31].

MobiSys'12, June 25–29, 2012, Low Wood Bay, Lake District, UK.
Copyright 2012 ACM 978-1-4503-1301-8/12/06 ...$10.00.

With cellular network data becoming more available, it is tempting to think that creating human mobility models from such data should be easy. This is, however, still not the case. For example, while CDRs readily yield insights into aggregate population densities, they do not convey whether their associated locations correspond to home, work, or other important places for particular cellphone users. Without such semantic information, it is difficult to abstract CDRs into models applicable to scenarios, regions, or populations that vary from those for which the real-life CDR data was collected. Furthermore, both the spatial and temporal granularity of CDR data is quite coarse. Spatially, CDRs are only accurate to the granularity of celltower spacings. Temporally, CDRs are only generated when phones are actively involved in a voice call or text message. Our work makes key contributions in overcoming the challenges stemming from lack of semantic information and coarse granularity, in order to produce usefully accurate models for arbitrary metropolitan regions.

Our modeling approach intelligently samples the spatial and temporal probability distributions from CDRs, or other population data, to generate sequences of locations and times for any number of synthetic people in any region for which the required distributions can be obtained. A generative model derived from CDRs has flexibility, compactness, and availability advantages over using CDRs directly. First, our models offer the option of perturbing the input distributions to evaluate what-if scenarios, for example to consider how the addition of a new residential or employment area might change traffic patterns. In contrast, the original CDRs are difficult to manipulate in meaningful ways. Second, our model for a metropolitan area with a 50-mile radius can be stored as a set of histograms that fit within 2 gigabytes. In contrast, an anonymized CDR dataset for the same area occupied approximately 100 gigabytes. Finally, our models can be made available to a larger research community because they do not to reproduce the mobility pattern of any individual real person. They thus avoid many of the privacy concerns associated with source CDRs.

The final stage of our modeling approach produces locations and times in the form of synthetic CDRs. These synthetic CDRs have the same format and call/text frequency characteristics of real CDRs. They are modeled to approximate the actual movement patterns of users. Increased model complexity results in more accurate movement patterns, which in turn produces higher-fidelity synthetic CDRs. We chose the CDR output format for several pragmatic reasons. One, we can compare this output directly against real CDRs, our best source of location information for large populations and regions. Two, this output can plug in directly into the growing body of analysis software that uses CDRs as input.

In this paper, we propose and evaluate WHERE ("Work and Home Extracted REgions"), a region-scale modeling approach. First, we identify the key properties of human movement, such as important locations and commute distances, that need to be represented as probability distributions. Then, we describe how these probability distributions can be used to generate synthetic CDRs for an arbitrary number of synthetic people.

We validate our approach by comparing the spatiotemporal dynamics of synthetic populations generated by WHERE to those of real populations. In particular, we compare the spatial population densities on an hourly basis for synthetic and real CDR sequences. Our validation begins with stylized examples that confirm our models' fidelity both quantitatively and visually. We validate both at the aggregate level, where simpler models may perform well, as well as at a finer granularity, which exposes the advantages of WHERE compared to other models considered. We then scale up our validation to large datasets containing real anonymized CDRs for the Los Angeles (LA) and New York City (NY) metropolitan areas. Our LA and NY datasets each span three months of activity for hundreds of thousands of phones, yielding billions of location samples.

Recognizing that real CDRs are not available to all researchers, we also evaluate models in which the same input distributions are derived from publicly available US Census data [32]. We show that models based on real CDRs closely approximate the real populations and movements of these cities. Models based on census data are also viable, but at a loss of significant accuracy.

Finally, we present example applications of our modeling approach. We create models for the LA and NY metropolitan areas and use the resulting synthetic CDRs to perform calculations that one may wish to perform on real CDRs. We show that calculations performed on the WHERE model produce far more accurate results than those performed on more naive models. For example, we can calculate daily ranges of travel that agree with real ranges, as well as perform more complex tasks such as investigating opportunistic message propagation in large urban environments.

The overall contributions of our work are:

- We introduce an approach to modeling human mobility patterns by generating fully synthetic CDRs from real-world probability distributions.
- Our approach works at the scale of large metropolitan areas and accounts for mobility differences between metropolitan areas.
- We show that our technique is extensible to greater levels of precision by providing it more complete input probability distributions (at the cost of increased model complexity).
- We validate our approach against large-scale location datasets drawn from two major US metropolitan areas. We compare our generated CDRs against real CDRs, and show that our location distributions achieve more than 4 times error reduction compared to a Random Waypoint model.
- As an example of how our models can help answer concrete questions about human mobility, we use our synthetic CDRs to compute daily ranges of travel. Our synthetic CDRs exhibit error at the median of less than 0.8 and 1 mile for NY and LA residents, respectively. This accuracy constitutes more than a 14 times improvement over that of a Weighted Random Waypoint model.

The rest of this paper is organized as follows. Section 2 presents the probability distributions needed to create our models. Section 3 explains the construction of synthetic CDRs. Section 4 describes our evaluation methodology. Section 5 validates our synthetic CDRs against stylized examples and Section 6 validates them against large datasets of real CDRs. Section 7 examines example uses of artificial CDRs. Finally, Section 8 discusses open issues and future directions, and Section 9 surveys related work.

2. Spatial and Temporal Parameters for Mobility Modeling

Modeling human mobility requires both spatial and temporal information about the places and times that humans move. Human mobility is tightly coupled to the geography of the city people live in [16, 17, 18, 25]. Therefore, any accurate mobility model should take into account both the area geography and individual user mobility patterns. In this section, we describe the probability distributions that serve as inputs to the WHERE method.

Figure 1 summarizes the overall flow of our approach, including key inputs and data structures. For each of the input distributions

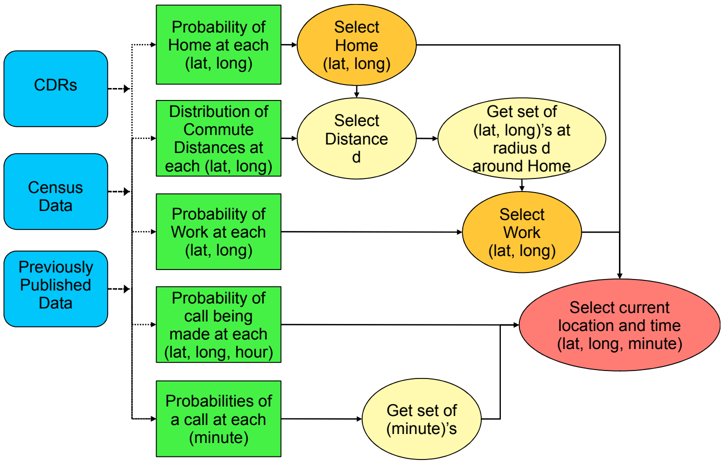

CDRs
Census Data
Previously Published Data
Probability of Home at each (lat, long)
Distribution of Commute Distances at each (lat, long)
Probability of Work at each (lat, long)
Probability of call being made at each (lat, long, hour)
Probabilities of a call at each (minute)
Select Home (lat, long)
Select Distance d
Get set of (lat, long)'s at radius d around Home
Select Work (lat, long)
Select current location and time (lat, long, minute)
Get set of (minute)'s

Figure 1: Overall view of the WHERE modeling approach. Five input probability distributions are important for our modeling technique, but their source can vary.

	Data Sources		
Distribution	All Public	Hybrid	All CDR
Home	Census	Census	CDR
CommuteDistance	Census	Census	CDR
Work	Census	Census	CDR
Hourly	Census Home and Work data split by time of day	CDR	CDR
CallTime and PerUserCallsPerDay	Previously-published work [1, 5]	CDR	CDR

Table 1: For the five probability distributions used in our approach, the input data for these probabilities can be gathered from different sources. We have explored methods based entirely on public data, as well as methods using data from proprietary anonymized CDRs.

required, Table 1 summarizes different methods for generating the distributions. The subsections below describe the distributions in more detail, and discuss the different possible sources for the input data required.

2.1 Spatial Information: Important Locations

Capturing where people spend time is important for creating accurate human mobility models. Prior work has demonstrated that the majority of people's movement occurs between a small number of locations. A full 60% of mobility can be accounted for with just the top two cellphone towers with which a user is associated [31]. Therefore, in modeling mobility, good accuracy can be expected if we have probability distributions for the few locations in which users spend the majority of their time.

For most people, the two most important locations are "home" and "work." A design decision for the model is how to gather input information and how to express the probabilities of different home and work locations. If one simply pulled locations independently from two separate home and work distributions, the accuracy of the resulting model would be quite poor, since such an approach would ignore the strong correlation between where someone works and where they choose to live. We also explored methods that assumed commute distances independent of home location, but these were also too inaccurate to be useful.

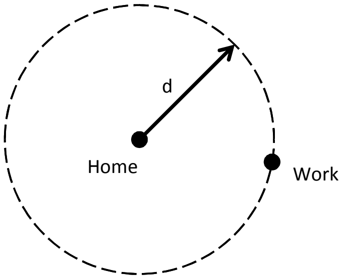
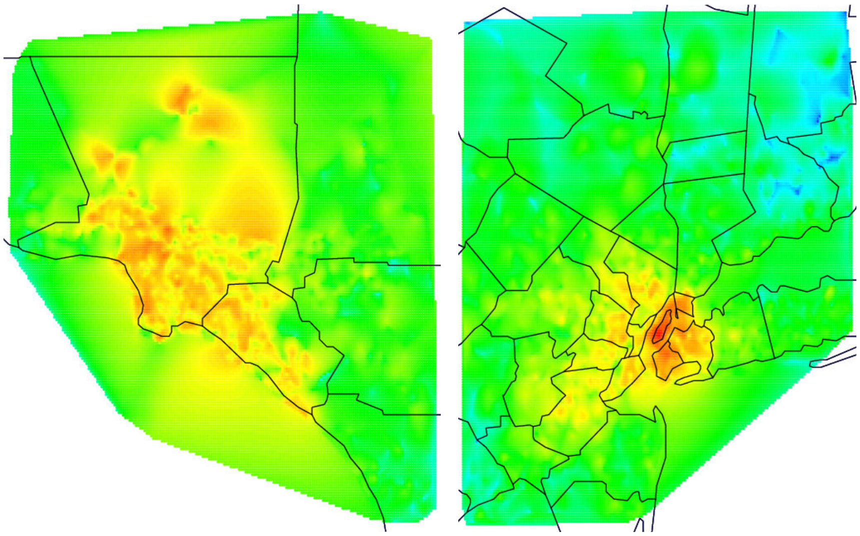


Figure 2: Selecting a user's home and work locations. Pick a home location, then select *d* from the distribution of commute distances at that home location. The work location is selected from a probability distribution for work locations on the resulting circle of radius *d*.

As shown in Figure 2, we use a different approach that estimates spatial home-work densities by relying on three distributions. First, we pull home locations randomly from a probability distribution across latitude and longitude expressing the likelihoods of where people live, i.e. *Home*. For each point in space, a second probability distribution *CommuteDistance* expresses the probability of having different commute distances, conditioned on that given home location. A commute distance, *d*, selected from this distribution can be envisioned as describing a circle of radius *d* around the selected home location. Next, our method selects a work location somewhere along this circle. To do so, a third distribution, *Work*, gives the probability of different work locations around a circle of commute distance *d* from the home location. These *Home* and *Work* locations are derived from population densities that correspond to the city we wish to synthesize. Such distributions could be computed from either real CDRs or the census data, which contains information about the number of people living or working in a particular area. We will discuss possible sources of each spatial distribution more thoroughly in a subsequent section.

2.2 Spatiotemporal Information: Hourly Population Densities

The distributions for *Home*, *Work*, and *CommuteDistance* give information about spatial probabilities, but do not link them to particular times of day. To improve the fidelity of our model, we need to include temporal information as well. For instance, a heavily



(a) Los Angeles

(b) New York

Figure 3: Logscale heatmaps of call densities in the LA and NY areas from 7 to 8 p.m. on weekdays over a 3-month period. (The apparent activity in offshore areas is only a byproduct of interpolation techniques used to produce these maps.)

residential area is likely to be more populated at night while a commercial district is likely to be more populated during the day.

We model the time-varying aspects of human mobility through the use of spatial population densities indexed by time t. Unlike the previous spatial distributions, which have specific meanings like home and work, this portion of our model is an aggregate distribution that simply reflects the probability of people being at a particular location at a particular time. For some it could be work, for others home, and for others neither.

Hourly population density distributions can be constructed in several ways. First, if available, CDR traces can be analyzed to estimate population at any point in space and time. This requires the assumption that the spatial densities of telephone calls is approximately equivalent to the spatial density of people. Prior work [17] has shown that cellular calls are an accurate representation of the locations of a user. In collections of actual CDRs, call locations are estimated as the location of the cell tower through which the call originated. Since cell towers vary in density (urban vs. rural) we interpolate the cell tower locations and call counts to a regular grid whose granularity is specified by the modeler. The hourly call counts per latitude-longitude grid area can be normalized to form a distribution across latitude and longitude called *Hourly*. Figure 3 shows instances of one-hour population density estimates made in this way for calling activity in the LA and NY regions.

If CDR data are not available, one could approximate the population densities from census data. For example, one could form an overall distribution by using census home location data to estimate population data during traditional non-work hours (e.g., 7pm to 7am) and census work location data to estimate population density during traditional work hours (e.g., 7am to 7pm). We evaluate this alternative in subsequent sections.

2.3 Temporal Information: Calling Patterns

The end goal of our modeling work is to create synthetic CDR traces that can be validated against real-world CDRs and that can be used in the same ways that previous research has used real-world CDRs. Thus, a final step in our model is to combine the spatial probabilities above with information about calling patterns, so that our model produces accurate synthetic CDRs for comparison and future use.

To incorporate realistic temporal information into the mobility model, we require realistic information about the distributions of user call rates. One can envision a user's daily call volume characteristics as being probabilistically chosen from a distribution of *PerUserCallsPerDay* indexed by different possible averages and standard deviation values.

Once a user's average and standard deviation of calls per day has been selected, the next question concerns the more detailed temporal patterns of when those calls are made. To this end, we separate users into similarity classes with separate temporal distribution functions for each similarity class. This method requires a detailed source of temporal input data, such as a real-world CDR trace. From this, we examine how many calls each user makes during each hour of the day, and this calling behavior gives us a 24-dimensional vector (i.e., one dimension for each hour). We normalize this vector so that it represents probabilities rather than call volumes, and then this normalized 24-dimensional vector is used to generate similarity scores between each pair of users. From this, we use a common clustering algorithm (X-means [27]) to cluster users. Given the option of one to 20 clusters of users, X-means determined 2 clusters of users to be the best choice for our data. Finally, we return to the CDR trace and record per-minute call probability distributions separately for each user class. From these we form a probability distribution *CallTime* for each UserClass and for each hour of the day. An example of empirically-collected call time probabilities is shown in Figure 4. While typical diurnal patterns are seen for both classes of users, one class clearly favors evening calls while the other favors afternoon calls.

Without access to CDR information, it is still possible to construct call distributions. One approach would be to assume that all users have similar temporal patterns, and simply use an overall

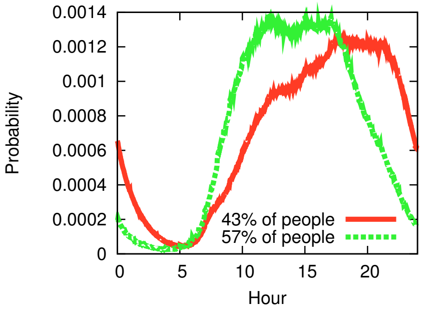


Figure 4: Distribution of call times for two classes of users as determined by X-means clustering.

Algorithm 1 Create

```
Ensure: $pop[]$ is an $N$ sized array of 4-element structures to be filled in with 4 properties for each of $N$ synthetic users
for $user = 0 \rightarrow N$ do
    $pop[user].home \leftarrow$ location from Home
    $commute \leftarrow$ distance from CommuteDistance conditioned on $pop[user].home$
    $pop[user].work \leftarrow$ location from Work at distance $commute$ from $pop[user].home$
    $pop[user].callsbehavior \leftarrow$ user type from CallTime
    $pop[user].callsperday \leftarrow \mu$ and $\sigma$ from PerUserCallsPerDay and independent of $pop[user].callsbehavior$
end for
```

per-hour call probability distribution function as the guide for when calls are most or least likely to be made. Such probabilities can be drawn from prior work including [1, 5]. Our work validates the accuracy of approaches like these against real-world CDR collections.

To summarize, the sources of temporal input data on call patterns can either be published statistics [1, 5] or proprietary CDR data. Because these call patterns have been seen multiple times in multiple contexts, we assume that such a calling pattern is general enough to hold regardless of the spatial data to which it is applied.

3. Algorithms for Model Generation

The previous section argued for the importance of particular spatial and temporal information as inputs towards a stochastic mobility model. In this section we describe how to use such information as inputs for mobility models with different degrees of complexity and accuracy. We start in Section 3.1 with a two-place model that has synthetic people who alternate between home and work. We call this model WHERE2. Despite the simplicity of this model, the attention to spatial and temporal distributions gives it considerable accuracy. Section 3.2 shows how the technique can be expanded into a three-place model we call WHERE3, with even more accurate results.

3.1 Two-Place Model: Work and Home

WHERE2 makes use of the fact that most people spend the majority of their time either at home or at work. The process of generating a synthetic CDR trace embodying this model occurs in two stages. In the first stage, a synthetic user is created according to Algorithm 1, *Create*. First, a user is assigned a home from *Home*. Second, a commute distance is selected for the synthetic user. As illustrated in Figure 2, *CommuteDistance* is conditioned on a user's particular home location. In this way, we begin to tie users specifically to the geography of the area to be synthesized. It is not sufficient to select from a general probability distribution for commute distances, because this may unfairly bias toward commute distances for very dense areas. Intuitively, the likely commute distances for a person living in midtown Manhattan are quite different from those who live in outlying exurbs. Third, all possible locations that are the selected distance away from the chosen home location are considered as possible "work" locations, and a work location is chosen for the synthetic user. Possible work locations are weighted with probabilities given by *Work*, and again are conditioned on a particular home location and commute distance.

Algorithm 2 Move

```
for $user = 0 \rightarrow N$ do
    for $day = 0 \rightarrow D$ do
        $callstoday \leftarrow$ normal random number with $\mu$ and $\sigma$ from $pop[user].callsperday$ distribution
        for $call = 0 \rightarrow callstoday$ do
            $calltime \leftarrow$ time from $pop[user].callsbehavior$
            $location \leftarrow$ location using probabilities of $pop[user].work$ and $pop[user].home$ at time $calltime$ from the Hourly
            print $user, day, calltime, location$
        end for
    end for
end for
```

Finally, after having selected a user's home and work, the synthetic user is assigned a calling pattern (i.e., mean μ and standard deviation σ calls per day) according to the distributions from *CallTime* and *PerUserCallsPerDay*.

In the second stage, our synthetic users are moved between "home" and "work" according to Algorithm 2, *Move*. Movement "occurs" based on synthetic CDRs representing calls made at different locations at different times. For each day, a number of calls to make is selected from the user's average (μ) and standard deviation (σ) of the number of calls made per day. Each call is then made according to the calling distribution of the user's class, as given in *CallTime*. When a "call" is made, the location of the call is determined to be either home or work according to the probability of a person being in the location at that time of day (i.e., the probability is determined according to the distributions in *Hourly*).

3.2 Extension to Additional Places

This section discusses how WHERE2 can be augmented for greater realism by increasing the number of important locations that are considered for each synthetic user. The tradeoff of model complexity against the fidelity of the synthetic trace can be tuned to suit the model's purpose.

We can extend WHERE using Algorithm 3, *CreateExpand*, to create WHERE3. When the synthetic user is given locations, a third location is selected such that the distance of the additional location is selected from a distribution of distances from the "home" and "work" locations of users in the real city, the *AverageDistance* Distribution. Clearly, this must be conditioned on the commute distance because distances between the three locations (home, work, and the one to be added) are conditionally dependent on each other (i.e., they form three sides of a triangle). Probabilities for the additional locations can be drawn from a probability distribution of the user locations over time, the *AllLocations* Distribution. Similarly, we can create WHERE4, WHERE5, etc. by incrementally adding locations to the previous model that are conditionally dependent on

Algorithm 3 CreateExpand

```
Ensure: $pop[]$ is an $N$ sized array of 5-element structures to be filled in with the 4 properties from $Create$ and 1 additional property for each of $N$ users
for $user = 0 \rightarrow N$ do
    Perform $Create$ for user
    $disttothirdloc \leftarrow$ distance of third cluster to $pop[user].home$ and $pop[user].work$ from AverageDistance
    $pop[user].third \leftarrow$ location from AllLocations at distance $disttothirdloc$ from $pop[user].home$ and $pop[user].work$
end for
```

the locations before them. Such models will be increasingly precise but come at the expense of making the models increasingly complex.

In addition, when the synthetic CDRs are created, probabilities for a user "moving" to a given location must be modeled realistically. In our approach, they are conditioned on the location of the home cluster. This needs to be done so that home and work remain the most frequently visited locations instead of a generically more popular third location. This way, additional locations will be visited in a way that mimics the way those locations are visited by real people. This requires modification of the hourly probability distribution described in Section 2.2, a step not reflected in the *CreateExpand* algorithm. Specifically, we need a set of hourly probabilities for each possible home location. By conditioning the probabilities for Section 2.2 by the location of the synthetic user's home, we greatly increase the size of the required set of input probability distributions, but we are able to add realism to human movement patterns. Once the probabilities are modified, no modification needs to be done to the *Move* algorithm except that when selecting from the hourly probability distribution, the home location must be taken into consideration as well.

4. Evaluation Methodology

This section offers methodology and background information regarding our experimental evaluations. In particular, we discuss the metrics by which we gauge model accuracy, the other mobility models we compare against, and potential sources for the input data our model requires.

4.1 Earth Mover's Distance

Ultimately, the synthetic CDRs are intended to create movement patterns that mimic those seen in the real CDRs. Thus, in aggregate, a "good" synthetic trace has the synthetic user population distributed in a very similar way in space as the real trace, at any point during the day. To quantitatively assess the similarity or dissimilarity of two population density patterns at a given time, we need a measure for comparing two spatial probability distributions (i.e., the hourly locations of real and synthetic users). To this end, we rely on the Earth Mover's Distance (EMD) as our measure of choice [29, 30]. To calculate it efficiently, we use the Fast EMD code from [26].

EMD attempts to find the minimum amount of energy required to transform one probability distribution into another. If one visualizes a probability distribution as a hill of earth to be reshaped into the second probability distribution, this energy is given by the "amount" of probability to be moved and the "distance" to move it. Since different distance weightings lead to different EMD values, we provide Table 2 as a basic calibration for the meaning of EMD values in this work. Table 2 provides EMD calibration data for a location shift transform. The starting point is a probability distribution in which all of the probability is concentrated in a single location (e.g., delta function). The "Location Shift" transform maintains the concentrated probability spike, but shifts it linearly by the specified number of miles to a new geographic location.

Miles	EMD
1	8.67e+05
10	8.65e+06
20	1.74e+07
30	2.61e+07
40	3.48e+07
50	4.35e+07

Table 2: EMD values for linear shifts of the probability distribution.

Table 2 confirms that in the case of a simple location shift, the EMD changes linearly with the shift amount or distance. In particular, an error caused by a simple shift of 1 mile corresponds to an EMD of 8.67e+05, and larger shifts scale proportionately. For the remainder of this paper we normalize our EMD errors by this factor. This allows us to refer to them in terms of the more human readable "miles of error". To determine the average error between real and synthetic location density patterns, we generate location probability distributions for each hour of the day from our synthetic CDRs. Then we calculate the EMD between our synthetic distributiond and the reference probability distribution (one of the distributions from Section 4.2). To convert to miles of error, we simply divide the result by 8.67e+05.

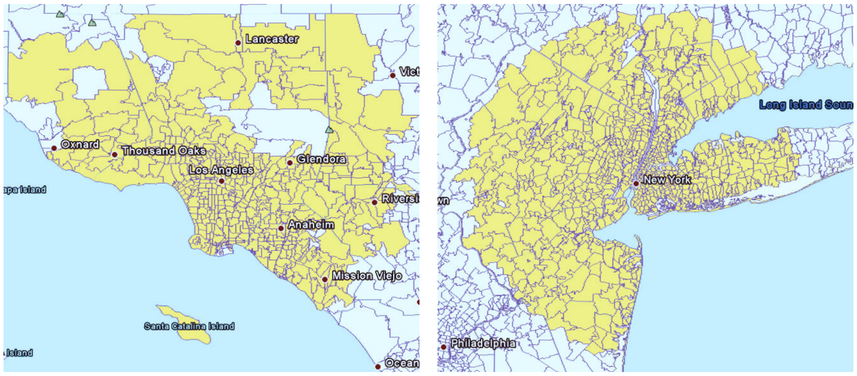
Another benefit that comes with using EMD in this context is that since we deal with probability distributions and physical distances, EMD is a metric in the strict mathematical sense [2]. The ensuing properties regarding relative distances (e.g. triangle inequality) are useful when comparing the fidelity of different models or use cases.

4.2 Comparison Models

To demonstrate the value of our approach, we compare our synthetic models against two commonly-used mobility models: Random Waypoint [19] and Weighted Random Waypoint.

Random Waypoint: In the Random Waypoint (RWP) model, each user selects a random destination from all possible destinations in the area to be simulated [19]. Once a destination is selected, the user moves at a random velocity toward the destination. When the selected destination is reached, the user waits for a random amount of time and then selects a new destination and new velocity to begin the procedure anew. While known to be simplistic, RWP has nonetheless been used extensively by the mobile computing research community, because so few alternatives exist.

Weighted Random Waypoint: The Weighted Random Waypoint (WRWP) model behaves similarly to RWP in that a user moves at a random velocity towards a chosen waypoint and waits for a random amount of time. However, in this variant, the destinations are not chosen from a uniform distribution. Instead, a location probability distribution is used to weight possible waypoints. The distribution we use is obtained by combining all of the hourly probabilities from Section 2.2 into a distribution that gives the popularity of the location for making calls over the whole day. This results in waypoints being preferentially chosen based on their popularity throughout the day. Intuitively, this accounts for the importance of home and work locations, so it is interesting to note how our approach further improves on this.



(a) Los Angeles (b) New York

Figure 5: ZIP codes in the LA and NY metropolitan areas used in our study. Note that NY and LA areas are drawn to the same scale.

	LA	NY
Total Unique Phones	318K	267K
Total Unique CDRs	1395M	1095M
Median CDRs/phone/day	18	18
Median calls/phone/day	6	7
Median texts/phone/day	6	5

Table 3: Characteristics of the CDR datasets for the LA and NY metropolitan areas. Each dataset spans 91 consecutive days from April 1 to June 30, 2011.

For either of the RWP and WRWP models, we generate synthetic agents that move on a grid and at random points in time, announce their location. In this way, the baseline models exhibit the same non-continuous behavior as the real CDRs. Once the baseline movement trace is generated, we calculate the EMD of a simulated hour of the baseline model against the reference CDRs.

4.3 Sources for Input Probability Distributions

As noted in Section 2, the probability distributions that form the input to WHERE can be obtained from a range of sources. We describe our sources here.

4.3.1 Real Call Detail Records

Call Detail Records maintained by cellular network operators provide up-to-date and low-cost information about human locations on a large scale. We have access to anonymized CDRs from a major US carrier, from which we extract the necessary spatial and temporal probability distributions.

Dataset Contents: We gathered location information for a random set of cellular phones whose billing addresses lie within the metropolitan regions of interest. First, we identified all ZIP codes within a 50-mile radius of the Los Angeles and New York city centers. These ZIP codes correspond to the darker colored regions in Figure 5. Second, we obtained anonymized and simplified CDRs for a random sample of phones registered to individuals with billing addresses in those ZIP codes. These CDRs contain the following information: a unique phone identifier in place of the telephone number, the starting time of the voice or text event, the duration of the event, and the locations of the starting and ending cell towers associated with the event. We excluded phones registered to businesses because their billing ZIP codes do not generally correspond to people's homes. We also excluded phones that appeared in their billing ZIP codes fewer than half the days they had voice or text activity, so as to exclude people who do not live in those ZIP codes.

Table 3 describes our CDR datasets for LA and NY. Each dataset contains more than a billion location samples for hundreds of thousands of phones over 3 months of activity, with 18 median location samples per day for each phone.

Data Validation: In previous work, we have verified that CDR datasets gathered using this same methodology accurately represent the mobility patterns of the population at large. More specifically, we have performed a number of comparisons against data from the US Census [32] and against ground truth provided by volunteers. One, we have confirmed that the number of sampled phones in each ZIP code is proportional to the population of that ZIP code [18]. Two, we have shown that the maximum pairwise distance between any two cell towers contacted by a phone in one day is a close approximation for how far the phone's owner traveled that day [17]. Three, we have shown that applying certain clustering and regression techniques produces accurate estimates of important locations in people's lives, in particular home and work [16].

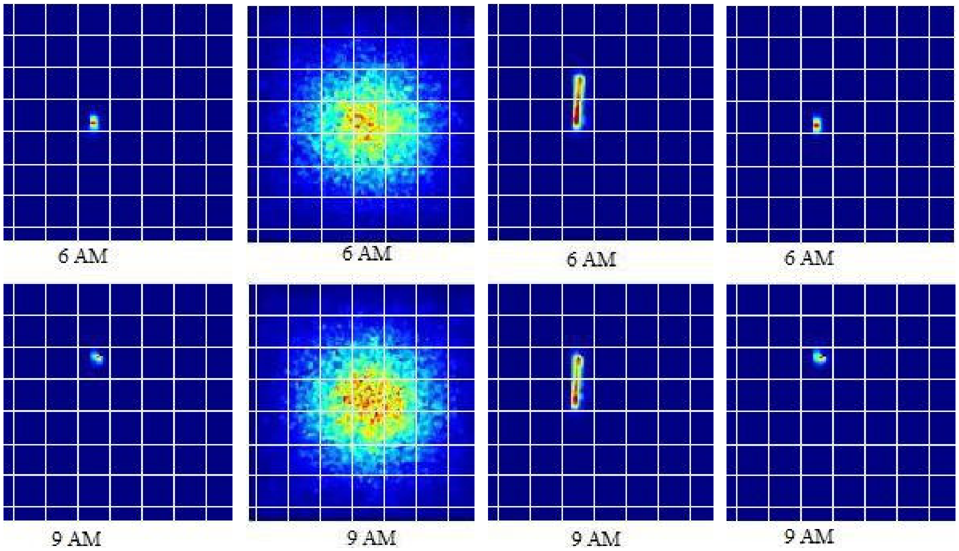
Privacy Measures: Given the sensitivity of location information, we took several steps to preserve the privacy of individuals represented in our datasets. First, only anonymous records were used in this study. Personally identifying characteristics were removed by someone not involved in the data analysis. Second, we worked only with large collections of phones. We did not focus our analysis on any individual phone and we present only aggregate results.

In addition to these active steps, it is in the nature of CDRs to yield only temporally sparse and spatially coarse location information. A CDR is generated only when a phone is used for a call or text message—at all other times the phone is invisible to us. Furthermore, we only know phone locations to the granularity of a cell tower. Because a tower often covers an area greater than one square mile, our spatial resolution is limited.

4.3.2 Census Data

Clearly, not all researchers will have access to real CDRs. However, in the US and other countries, there is publicly available data regarding home and work locations, as well as commute distances, for large populations. To demonstrate the efficacy of using publicly available data, we used census information for the same regions defined in Section 4.3.1 to construct probability distributions for the locations of home and work.

Although the census can be used to determine densities of home and work locations, it provides little or no information about the hourly probabilities of a given location. Therefore, when we use WHERE on All Public data, we must make an assumption about the hourly distributions. To that end, we assume that the *Work* distri-



(a) Two Locations Ideal (b) RWP (c) WRWP (d) WHERE2

Figure 6: Heatmaps comparing the probability distributions of three models of mobility against the "Two Locations" test case during the "home" and "work" phase of the test case for 10,000 synthesized users. Hot (red/light) locations have more calls made at that time, while cold (blue/dark) locations have fewer.

bution doubles as the hourly distribution of all hours between 7am and 7pm on Weekdays and that the *Home* distribution provides the hourly distribution at all other times.

We also consider a Hybrid scheme in which we use a combination of public data and CDRs. In this scheme, home, work, and commute distances come from the census, but the other distributions are drawn from CDR data.

4.4 Artificial Test Cases

It is useful to construct some stylized examples to demonstrate the function of specific features of WHERE. By creating artificial test cases, we are better able to reason about the expected behavior of the model and its strengths and weaknesses. In this section, we explain how we construct three artificial reference probability distributions that exercise various features of WHERE.

Two Locations: Initially, we must determine that WHERE functions correctly. The model must be able to place synthetic users at home and work locations and move them according to time of day. Further, since we emulate CDRs with discrete call locations, the synthetic users must only exist at these locations.

We test these features with the "Two Locations" test case. The entire world is populated by users that move predictably between two locations at highly regimented intervals. From 7am to 7pm on weekdays, all of the probability is concentrated in a single "work" location. At all other times, the probability is clustered in a second "home" location.

Two Distinct Types: We then expand the test cases to include multiple possible home and work locations. WHERE must be able to correctly condition commute distances on home locations. If multiple possible works exist for a home, the model must select a realistic one.

We demonstrate this with the "Distinct Types" test case. The home distribution has half the probability at one location and half at a second location. The work distribution is also split between two locations. The commute distributions are constructed such that each home location uniquely identifies a work location.

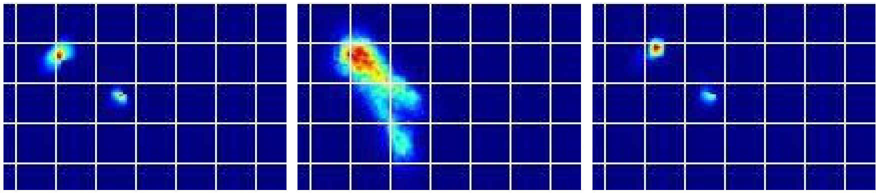
5. Validation: Stylized Examples

In this section we evaluate the accuracy of WHERE2 in synthesizing roughly 10,000 people using input probability distributions for the stylized test cases described in Section 4.4. We compare, both graphically and quantitatively, errors in the WHERE2, RWP, and WRWP models relative to an ideal result for each stylized case.

5.1 Two Locations Test Case

Figure 6 illustrates the spatial density patterns of user locations created by the "Two Locations" test case. Figure 6(a) shows the ideal result, in which probability is a spike at the home location at 6am, and a spike at the work location at 9am. RWP (Figure 6(b)) clearly differs greatly, because it is given so little input information regarding spatial or temporal patterns to model. The WRWP model better concentrates probability near the home and work locations, but does not have sufficient temporal input to distinguish how 9am mobility probabilities should differ from the 6am ones. Finally, Figure 6(d) shows the heat maps for WHERE2, which are a strong visual match to the ideal case. Even though WHERE2 has no hardwired information about what "work hours" are, it is able to correctly model the call distribution information both during the "home" and "work" phases. In contrast, while the WRWP model is able to correctly move users from "home" to "work", the lack of timing information damages its functionality for generating realistic synthetic CDRs.

While Figure 6 allows for an easy visual differentiation between the different models, the EMD measure is better for quantifying the observed differences. In fact, the EMD for WHERE2 corresponds to a simple shift of the cumulative probability of less than 0.5 miles. This is a 10X improvement over WRWP, the next best model considered. The WHERE2 model correctly generates calls from the desired discrete locations, and among the considered models, it is the only one to recognize that not all locations are equally probable at all times. Because each user only has two locations, WHERE3 performs exactly like WHERE2.



(a) Two Distinct Types Ideal (b) WRWP (c) WHERE2

Figure 7: Heatmaps comparing the WRWP model to WHERE2 for 10,000 synthesized users. The WRWP is unable to account for the two distinct patterns and thus sends synthetic users on paths never taken in the "real" trace.

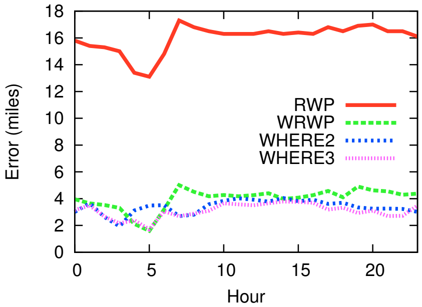


Figure 8: EMD error over time for different models of human movement in the NY area. Our expanded model shows average errors smaller than 3 miles. WHERE fits between WHERE3 and WRWP.

5.2 Two Distinct Types Test Case

Figure 7 shows the heatmaps of the "Two Distinct Types' ideal distribution, as well as WRWP and WHERE2. WHERE2 detects that users make calls only from one of two locations and thus correctly positions users. In contrast, the weighted model has a single user travel to each of the four possible locations, resulting in a significantly worse EMD. When comparing against the WRWP model, WHERE2 has an average of 5% improvement across all hours of the day. Because each user type only has two locations, WHERE3 again performs exactly like WHERE2.

6. Validation: Large-Scale Real Data

In this section, we evaluate the accuracy of WHERE in synthesizing more than 10,000 people using probability distributions from the datasets of real CDRs described in Section 4.3.1, and from the census data described in Section 4.3.2. We compare, both graphically and quantitatively, errors produced by the WHERE, RWP, and WRWP models relative to those two large-scale real-data sources.

6.1 Modeling Based on Real CDRs

Figure 8 shows EMD errors for the different possible models using distributions drawn from real CDRs. By including population density information, WRWP improves over original RWP by 4 times. WHERE, however, outperforms this WRWP model by an additional 20%. Furthermore, WHERE3 improves further still; it is able to recreate the NY probability with an accuracy of about 3 miles. WHERE and WHERE3, therefore, provide a powerful tool for obtaining accurate large-scale motion patterns that result from individual user movements.

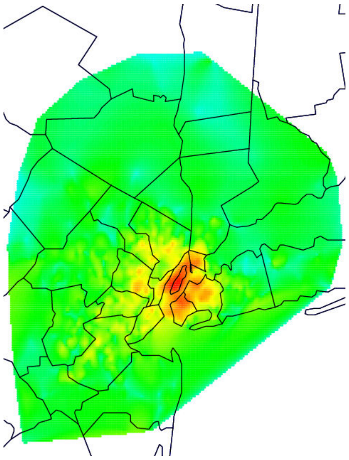


Figure 9: Logscale heatmap of the output of WHERE for the NY area.

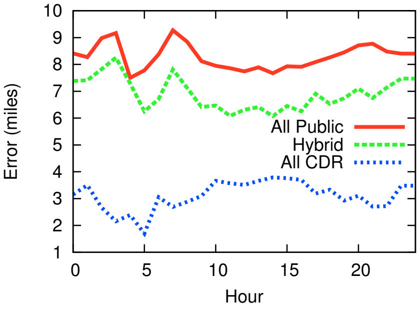


Figure 10: EMD error using different levels of available data. Even with fully public data, errors remain at an average of 8 miles.

Finally, Figure 9 gives an overall visual demonstration of WHERE's effectiveness. When this heatmap is compared to Figure 3(b), the two distributions are found to be visually very similar. The low EMD error combined with the visual confirmation of similarity shows that WHERE is an effective approach to modeling human mobility in a city.

6.2 Modeling Based on Census Data

Although the most accurate synthetic CDRs are derived from real CDRs, other data sources can be used at some loss in accuracy. For example, Figure 10 shows that using all-public data from the US Census, we are able to recreate location distributions in NY with an average error of 8 miles. Using a hybrid of census data with some CDR information regarding the call densities at different times of the day (not publicly available, but easily aggregated and thus potentially easier to obtain) reduces this error to an average of 6.8 miles. In contrast, using all-CDR information reduces the average error to some 3 miles in WHERE3, as shown earlier.

Although 8 miles of error seems large compared to the 3 miles of error possible with WHERE3, this error is aggregated across all users. Individually, user behavior remains qualitatively correct and

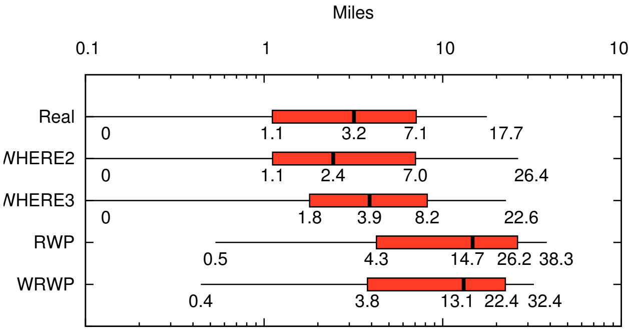


Figure 11: Daily Range from the real NY data as compared to WHERE and two variants of Random Waypoint. WHERE shows more realistic behavior than either RWP or WRWP.

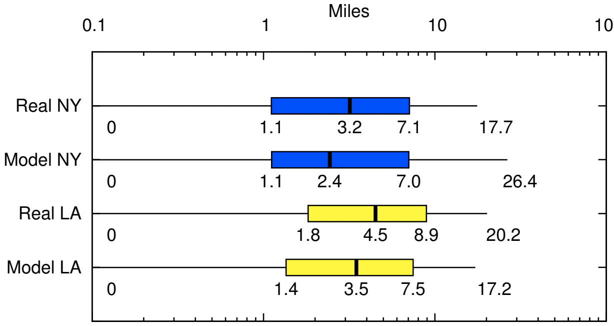


Figure 12: Comparing daily range statistics for NY and LA generated by WHERE2 and with real data. WHERE2 accurately recreates the characteristics of NY and LA mobility.

relative differences remain correct. Thus, though one could not rely on fully public data for an application such as placement of cellular towers, there remain many uses at such error levels. Notable examples include determining commute distances or experimenting with hypothetical population shifts (see Section 7.1 and Section 7.3), both of which depend on relative user behavior.

7. Example Uses

Creating synthetic CDRs has a wide range of implications for the scientific community. The inability to access real CDRs has frustrated many researchers. In generating artificial, yet realistic CDRs, we enable the greater community to perform a wide range of experiments, with an assurance that the results mimic those that would have been obtained with real CDRs.

This section presents three scenarios that highlight the value of our mobility models based on synthetic CDRs. We first present two usage experiments that we performed on synthetic CDR traces as well as on real CDRs. In Section 7.1 we replicate the daily range calculations reported in [18], which defined daily range as the maximum distance a person travels in one day. This comparison serves as an important, independent validation of WHERE because no daily range statistics were included in the input to our models. By performing well on the daily range metric, WHERE demonstrates an ability to model large-scale movement patterns while retaining realistic individual behavior. Second, in Section 7.2, we perform a simple message flooding experiment relevant in the context of opportunistic networking. Our third example highlights the major benefit of synthetic CDR traces in their ability to predict and visualize the impact of hypothetical changes to regional mobility patterns. Such changes might result, for example, from new employers moving into the region or new residential areas being developed. With this in mind, Section 7.3 presents an example of using WHERE to generate synthetic CDRs for a what-if scenario regarding mobility in the New York region.

The scenarios presented in this section constitute a set of example applications that demonstrate the broad utility of our models. These examples are far from exhaustive, but are important as illustrations of a wide range of possible uses.

7.1 Daily Range

Daily range, or the "diameter of a convex hull," has been shown to be a useful tool for characterizing human mobility patterns [18, 20]. We can therefore demonstrate the value of our modeling techniques by showing that synthetic CDRs they generate can be used to replicate daily range experiments and match the results obtained with real CDRs.

Additionally, daily range serves as an important metric for verifying correctness of the generated models. The EMD metric measures the aggregate behavior of the users, but daily range displays the results at a per-user granularity. Correctness at a metropolitan scale may or may not be correlated with correctness at the scale of individual users. The daily range metric tests whether individuals are modeled realistically.

We summarize our results with the help of boxplots in Figure 11. The boxplots depict five-number summaries of the complete empirical distributions of interest. The "box" represents the 25^{th}, 50^{th}, and 75^{th} percentiles, while the "whiskers" indicate the 2^{nd} and 98^{th} percentiles. The horizontal axes show miles on a logarithmic scale. Nearly any difference between the medians is statistically significant due to the large sample sizes.

Figure 11 shows the daily range results for synthetic CDRs produced by WHERE2 and WHERE3, as well as those produced by RWP and WRWP. It is clear that the behavior of the random waypoint models differs greatly from reality. For example, WRWP exhibits qualitatively wrong results and errors greater than 300%. In contrast, WHERE2 is qualitatively correct, and its median daily range value is within 0.8 miles of the true value. WHERE3 closes this error further: 0.7 miles error at the median.

In other words, this daily range example shows that WHERE2 and WHERE3 capture aspects of mobility that are not captured by the random waypoint models. By exposing differences at the grain of individual synthetic users in this way, WHERE2 results in more than a 14X improvement over WRWP. This advantage was not visible under the more aggregate metric of EMD, when WHERE appeared to hold only a small advantage over WRWP.

This example also demonstrates another attractive feature of our model. Because much of human mobility is between two major points of interest, adding a third point to our model provides very little obvious benefit for experiments of this nature. A researcher interested in problems such as daily range can choose to use the simpler WHERE2 model, avoiding the effort of computing the more complex WHERE3 model with little or no loss of accuracy.

Comparing NY and LA: Though much of the evaluation thus far has focused on New York City, the technique is more broadly applicable. Figure 12 compares daily range statistics for both NY and LA, for both real CDR data as well as WHERE. As before WHERE generates very high-accuracy results for these statistical

	Real	WHERE2	WHERE3
Delivery Percentage	98%	53%	83%
Median Message Delay	17 hours	74 hours	18 hours

Table 4: Message passing properties for opportunistic flooding. For such a scenario an expanded model of human movement gives better results.

distributions. At the median value, WHERE computes daily range for the LA area with only 1 mile of error. Low errors are seen at the other percentiles as well. This demonstrates WHERE's applicability across cities with very different mobility patterns and geographic characteristics.

7.2 Message Propagation

Another use for CDRs is the investigation of human mobility for social contacts, epidemiology, and data carrying. For instance, in a delay-tolerant networking scenario [8], what sorts of routing algorithms work best? In this section we demonstrate that our artificial CDRs can be used to simulate inter-person contacts.

To that end, we developed a simulator for epidemic routing [33]. At random times in the CDR trace, a message is injected from a random source to a random destination. As users meet they exchange all messages. For the purposes of this simplified experiment, users "meet" if their last known positions are the same within one hundredth of a degree of latitude and longitude (roughly a circle with a radius of 0.5 miles). When the message reaches its intended recipient, the time is recorded. This allows us to examine properties of the CDRs, including delivery percentages and message delay rates. Note that only a small subset of all users is considered for this experiment, as our full user base would ensure message delivery regardless of movement patterns. Here, we limit the experiment to 5,000 messages passed between 5,000 users in the New York area over 30 days.

Table 4 displays statistics regarding message passing success in the real CDRs as well as in our synthetically constructed models. Although even WHERE3 is far from the actual values, it is clear that adding the third location provides significant improvement. This is consistent with the idea that opportunistic networking relies on highly mobile users [36].

We stress that these results are based on the most basic version of the model. This example is provided to demonstrate the extensibility of the model as more clusters are added.

7.3 Hypothetical Cities

A major advantage of mobility models over empirically-gathered real-world CDRs is the ability to create parameterized model of cities and user behavior patterns that cannot be observed in the real world. This allows researchers and city planners considerable power to experiment with the effects of modifications that are being considered.

As one example, Figure 13 shows a box plot of data that can be constructed from artificial CDRs if we imagine a scenario intended to study the impact of telecommuting. In this case, we adjusted the WHERE2 model for NY so that a randomly selected 10% of the people whose original work locations were in the borough of Manhattan were instead given work locations that matched their home location. Such a scenario might arise, for example, if tolls or subway fares were increased, or if the city enacted policies to encourage telecommuting. With only 10% of the Manhattan workers affected, the impact on daily range is subtle, but noticeable; the results do demonstrate the trend that daily ranges would be slightly diminished by this change. In general, the power of such examples is in allowing model users to explore the impact of different mobility and behavioral changes on their metric of interest.

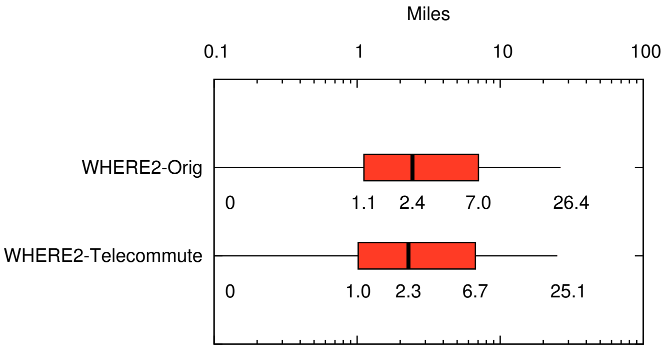


Figure 13: Boxplots of WHERE applied to a NY scenario in which 10% of Manhattan workers opt to telecommute.

This example demonstrates the flexibility of our modeling approach. Experiments that are not feasible in reality (e.g., convincing 10% of New Yorkers to telecommute) are simply a matter of adjusting some input probabilities.

8. Discussion and Future Work

As we have shown, our approach to human mobility modeling achieves its three main goals: capturing the motion of individuals among important places in their lives, aggregating that motion to reproduce human densities over time at the scale of a metropolitan area, and accounting for differences between metropolitan areas. However, there remain areas for refinement.

8.1 Travel Routes and Additional Locations

One such refinement would be to produce not only sequences of locations with associated times, but also routes taken between those locations. A tradeoff we incurred in obtaining location information for large populations in wide geographic regions is that both real CDRs and census tables are too coarse to provide route information. So far we have not attempted to improve on the spatial and temporal granularity of our source data, although other work [4] has had some success in identifying routes by making use of individual cellular antennas (not only towers that hold multiple antennas) and by analyzing longer sequences of antennas involved in the same call (not only the start and end towers).

In future work, we plan to infer such routes by using maps of transportation networks to interpolate between the locations produced by our current models. A similar idea has been applied to finer-grained location traces in smaller geographic regions, in particular traces of WiFi access-point associations in a university campus [34]. Working at the scale of metropolitan areas adds complexity because it becomes important to distinguish between modes of transportation beyond walking, e.g., driving a car, taking a train, or riding a bicycle. We can obtain some of the necessary information from public sources such as Table P30, Summary File 3, from the 2000 US Census: "Means of Transportation to Work for Workers 16+ Years" [32]. In addition, we can apply heuristics based on the distance and time between two location samples.

Once we select a mode of transportation between two locations, we can run proven routing algorithms such as those used by popular websites that provide driving, public transportation, biking, and walking directions [14]. Human mobility models that include route estimates would have broader applicability than those that produce only locations and times.

In addition, WHERE currently restricts users to a limited number of places. Future research directions may include incorporating heuristics about additional locations that a person may visit, perhaps by using WRWP as a rare, but possible additional destination. While it is true that the majority of movement occurs between important locations, there may be applications (such as the message passing experiment listed in Section 7.2) that could benefit from information about the long tail of other locations a person visits.

8.2 Differential Privacy

It is also worthwhile to discuss why we believe that our approach to human mobility modeling will preserve privacy. The intuitive reason is that we are careful not to reproduce the mobility pattern of any individual real person, for example someone represented in an input dataset of real CDRs. Instead, we create synthetic mobility patterns by sampling a sequence of probability distributions that each represents a large population. This approach makes it highly unlikely that any of our synthetic people exhibits a mobility pattern that identifies a single real person. However, the approach does not ensure that an adversary will be unable to reverse our algorithm to arrive at some portion of the source dataset, especially an adversary who brings to bear auxiliary datasets.

In future work, we plan to formalize our privacy argument by adjusting our algorithm to achieve *differential privacy* [7]. Informally, a differentially private algorithm is one that produces approximately the same output on two input datasets that differ only in the data for one individual. Our modeling approach naturally lends itself to a proven technique for achieving differential privacy without significantly affecting accuracy, namely to introduce controlled noise at key points in the algorithm. For example, we could introduce noise in the input CDRs, or when sampling one or more of our input probability distributions. Similar techniques have been successfully applied to network traffic traces [23].

We plan to show that our adjusted algorithm has the property that an individual's presence or absence in the input dataset will not alter the output by a significant amount. This property provides strong privacy guarantees in an information theoretic sense, regardless of how much auxiliary data an adversary applies.

9. Related Work

Characterizing human mobility based on cellular network data has recently received considerable attention. In our previous work [16], we developed an algorithm for identifying people's important locations based on anonymized cellular network data, and showed how to use these important locations to estimate home-to-work commute distances and commute carbon footprints for large populations in the New York and Los Angeles metropolitan areas. In other previous work [18, 17], we also characterized the daily range of travel of those same populations. Girardin et al. used cell phone usage within cities to determine locations of users in Rome [9], New York City [11], and Florence [10]. They were able to find where people clustered in these cities and the major paths people tended to take. The work presented in this paper goes beyond characterization to develop algorithms for constructing human mobility models from both cellular network data and census data.

A number of previous efforts attempted to model human mobility at various scales. Rhee et al. [28] studied statistical patterns of 44 participants carrying GPS devices for four months and concluded that people's movement has a resemblance to Levy flights, random walks where the step-length probability has a heavy-tailed distribution. They also proposed a Levy walk mobility model that can be used for network simulations. Kim et al. [20] developed an algorithm for extracting a human mobility model from wireless network traces collected from WiFi APs at Dartmouth College. Yoon et al. [34] present a trace-driven framework for generating realistic mobility models based on the association information between WiFi users and access points and maps of the area where WiFi traces were collected. Hsu et al. [15] used WLAN traces to extract a time-variant community mobility model that showed good performance via simulation. In contrast, we developed and validated mobility models based on CDRs for hundreds of thousands of people moving across large metropolitan areas.

González et al. [13] used cellular network data from an unnamed European country to create a universal model of how individuals move. Song et al. [31] studied similar cellular network data to predict an individual's movements. Specifically, the authors consider the towers associated with phone users and show that given sufficient past history, one could guess the current location of a given user with high accuracy. Other efforts have developed algorithms for predicting where a user will travel next [3, 6, 22]. To our knowledge, we are the first to develop and validate models that account for differences between geographic areas.

Several recent papers have looked at the privacy risks of releasing location traces. Krumm [21] showed that home addresses of individuals who shared their GPS traces could be accurately identified based on the combination of reverse geo-coding and white pages. Golle and Partridge [12] used the publicly available census data to show that knowing a person's home and work locations, even at a coarse resolution, is often enough to uniquely identify that person. Finally, Zang and Bolot [35] showed that releasing anonymized CDRs in its original format poses serious privacy risks, as a large fraction of the population could be re-identified from the anonymous data. These papers motivated us to look beyond a simple anonymization of location traces. In the future, we plan to show that WHERE preserves differential privacy.

10. Conclusions

Modeling human mobility is important for mobile computing research, urban planning, epidemiology, and ecology. In this paper, we proposed a method for generating realistic human mobility models for large metropolitan areas from real Call Detail Records (CDRs) and census data. Our models generate spatio-temporal data in the form of synthetic CDRs, which could then be processed by the same algorithms that operate on the real ones. We also demonstrated how our method can be extended to create more precise models of human mobility.

We validated our methodology by first showing that our synthetic CDRs maintain key properties of the real CDRs, in particular human densities over time. We then demonstrated that our WHERE model is more accurate than other models such as Random Waypoint and Weighted Random Waypoint. Finally, we gave examples of experiments enabled by our models, and demonstrated that our synthetic CDRs perform well compared to the real CDRs. Specifically, we showed that we can calculate median daily ranges with an error at the median of less than 1 mile, and good agreement across the rest of the distribution. In future work, we plan to show that our modeling approach preserves differential privacy, at which point we hope to release our models to the broader research community.

11. Acknowledgments

We thank our shepherd, Rajesh Balan, and the anonymous reviewers for their feedback. Parts of this work were supported by the National Science Foundation under Grant Nos. CNS- 0614949, CNS-0627650, and CNS-0916246. Isaacman also acknowledges support from a Princeton University Wallace Memorial Fellowship in Engineering and a research internship from AT&T Labs.

12. References

[1] S. Almeida, J. Queijo, and L. Correia. Spatial and temporal traffic distribution models for gsm. In *Vehicular Technology Conference*, Sept. 1999.

[2] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek. Unsupervised clustering of multidimensional distributions using earth mover distance. In *Proc. KDD'11*, 2011.

[3] M. A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. *World of Wireless, Mobile and Multimedia Networks and Workshops*, 2009.

[4] R. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In *13th International Conference on Ubiquitous Computing (Ubicomp)*, Sept. 2011.

[5] J. Candia, M. C. González, P. Wang, T. Schoenharl, G.Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *MATH.THEOR.*, 41:224015, 2008.

[6] K. Dufková, J.-Y. Le Boudec, L. Kencl, and M. Bjelica. Predicting user-cell association in cellular networks from tracked data. *Intl. workshop on Mobile Entity Localization and Tracking in GPS-less Environnments*, 2009.

[7] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation (TAMC)*. Springer Verlag, April 2009.

[8] K. Fall. A delay-tolerant network architecture for challenged internets. In *SIGCOMM*, 2003.

[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, F. Dal Fiore, J. Blat, and C. Ratti. Understanding of tourist dynamics from explicitly disclosed location information. *Symposium on LBS and Telecartography*, 2007.

[11] 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 *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

[12] P. Golle and K. Partridge. On the anonymity of home/work location pairs. In *Seventh International Conference on Pervasive Computing (Pervasive 2009)*, 2009.

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

[14] Google Maps. `http://www.census.gov`.

[15] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy. Modeling time-variant user mobility in wireless mobile networks. In *IEEE INFOCOM*, 2007.

[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 *9th International Conf. on Pervasive Computing*, 2011.

[17] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in los angeles and new york. In *Eighth IEEE Workshop on Managing Ubiquitous Communications and Services*, 2011.

[18] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2010.

[19] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. In *Mobile Computing*, pages 153–181. Kluwer Academic Publishers, 1996.

[20] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IEEE INFOCOM*, 2006.

[21] J. Krumm. Inference attacks on location tracks. In *Fifth International Conference on Pervasive Computing (Pervasive 2007)*, 2007.

[22] K. Laasonen. *Mining Cell Transition Data*. PhD thesis, University of Helsinki, Finland, 2009.

[23] F. McSherry and R. Mahajan. Differentially-private network trace analysis. In *Proc. ACM SIGCOMM*, 2010.

[24] W. Navidi and T. Camp. Stationary distributions for random waypoint models. *IEEE Transactions on Mobile Computing*, 3(1):99–108, 2004.

[25] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PLoS ONE*, 2012.

[26] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.

[27] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *17th International Conf. on Machine Learning*, 2000.

[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 *IEEE INFOCOM*, 2008.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proc. IEEE International Conference on Computer Vision*, 1998.

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

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

[32] US Census Bureau. `http://www.census.gov`.

[33] A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. Technical Report CS-200006, Duke University, 2000.

[34] J. Yoon, B. Noble, M. Liu, and M. Kim. Building realistic mobility models from coarse-grain traces. In *Proc. ACM MobiSys*, 2006.

[35] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Seventeenth Annual International Conference on Mobile Computing and Networking (MobiCom 2011)*, 2011.

[36] G. Zyba, G. M. Voelker, S. Ioannidis, and C. Diot. Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd. In *IEEE INFOCOMM*, 2011.