From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data

Ilias Leontiadis
Telefonica Research
ilias@tid.es

Antonio Lima
University of Birmingham
a.lima@cs.bham.ac.uk

Haewoon Kwak
Qatar Computing Research Institute *
hkwak@qf.org.qa

Rade Stanojevic
Telefonica Research
rade@tid.es

David Wetherall
University of Washington
djw@cs.washington.edu

Konstantina Papagiannaki
Telefonica Research
dina@tid.es

ABSTRACT

Through their normal operation, cellular networks are a repository of continuous location information from their subscribed devices. Such information, however, comes at a coarse granularity both in terms of space, as well as time. For otherwise inactive devices, location information can be obtained at the granularity of the associated cellular sector, and at infrequent points in time, that are sensitive to the structure of the network itself, and the level of mobility of the device. In this paper, we are asking the question of whether such sparse information can help to identify the paths followed by mobile connected devices throughout the day. If such a task is possible, then we would not only enable continuous mobility path estimation for smartphones, but also for the millions of future connected "things".

The challenge we face is that cellular data has one to two orders of magnitude less spatial and temporal resolution than typical GPS traces. Our contribution is to devise path segmentation, de-noising, and inference procedures to estimate the device stationary location, as well as its mobility path between stationary positions. We call our technique *Cell**. We complement the lack of spatio-temporal granularity with information on the cellular network topology, and GIS (Geographic Information System).

We collect more than 3,000 mobility trajectories over 8 months and show that *Cell** achieves a median error of $230m$ for the stationary location estimation, while mobility paths are estimated with a median accuracy of $70m$. We show that mobility path accuracy improves with its length and speed, and counter to our intuition, accuracy appears to improve in suburban areas. *Cell** is the first technology, we are aware of, that allows location services for the new generation of connected mobile devices, that may feature no GPS, due to cost, size, or battery constraints.

*This work has been completed while the author was working at Telefonica Research

CoNEXT'14, December 2–5, 2014, Sydney, Australia.

ACM 978-1-4503-3279-8/14/12 ...$15.00.
http://dx.doi.org/10.1145/2674005.2674982 .

Categories and Subject Descriptors

C2.1 [**COMPUTER-COMMUNICATION NETWORKS**]: Wireless communication—*Mobility*; C2.3 [**COMPUTER-COMMUNICATION NETWORKS**]: Network monitoring

General Terms

Algorithms

Keywords

Mobility Modeling, CDRs, Network Events, Trajectory Estimation, Localization, Cellular Networks, Street Routing

1. INTRODUCTION

Mobile devices are increasingly being seen as a unique sensor to individuals and humanity as a whole. Pocket-sized devices tend to follow us throughout our lives, capturing our day to day routines and providing us with valuable services, that exploit our context and preferences. A significant amount of research has gone into developing technologies that allow such devices to enable novel computing interactions, based on an increasingly greater number of sensors on board. The fundamental obstacle that needs to then be overcome is that of limited battery power. During the past 5-10 years, the research community has focused on ways to improve battery consumption for smartphones and define techniques for energy efficient location acquisition, through sophisticated energy efficient sensor hubs, WiFi, or the careful duty cycling of GPS; see [17] and references therein.

While smartphones are reaching nearly ubiquitous penetration in developed regions, a whole new array of connected devices emerges and is projected to reach billions by the year 2020. These devices tend to be connected to the cellular network (through 2G/3G or 4G), and are oftentimes mobile (see sim-watch.com or the array of devices supporting fleet management). However, due to cost, size or battery constraints, rarely do such devices feature GPS capabilities or sophisticated sensor architectures. The question we are addressing with this work is whether there is a way of enabling this whole new set of devices with precise location path information by utilizing the fact that these devices are continuously connected to the cellular network.

Mobility management is a fundamental function of the cellular infrastructure. However, unlike recent public belief, *in normal op-*

eration a cellular network is not in a position to accurately localize every single subscriber in space. In most cellular infrastructures today, mobile subscribers leave a trace of their associated tower, only when making/receiving calls, sending/receiving SMS, crossing Location Area Codes (LACs) or are routinely polled by the network due to stale information that exceeds a preconfigured duration (on the order of a few hours). Full handover information [1] is further accessible through RNC-level monitoring. In this paper we approach the problem of localizing subscribers from such sparse and noisy data available to the cellular operator.

We develop *Cell**, an algorithm that is able to parse the continuous sector observations and identify the segments corresponding to a stationary position, and the segments corresponding to actual mobility. We use the stationary segments, that typically feature multiple sector observations, to derive more accurate estimations for the stationary position. Using those enhanced locations that go beyond the sector coverage area, we develop an algorithm that is able to extract mobility paths between stationary points. Our path estimation algorithm uses cellular topology information to identify the areas on the map that are consistent with the observed sector associations, and weighs them according to the likelihood for a mobile subscriber to have connected with a tower from different locations within the respective sector coverage area. It then biases the route selection to pass through high probability areas while respecting the underlying road network.

To understand the accuracy and the limitations of *Cell** we study continuous mobility patterns for more than 30 individuals, while recording their ground truth GPS locations, and their associated cellular sectors (for every single handover). Our measurements cover more than 3000 unique trajectories in a European country.

*Cell** is able to identify the stationary position of the connected device with a median error of 230 meters which is comparable to the GSM triangulation technologies that actively probe the subscribers' signal strength [22]. In case of mobile paths, *Cell** has even lower median error of 70 meters due to the de-noising and GIS information that enrich the path inference. Somewhat intuitively, we show that the accuracy of our algorithm improves with the speed of the connected device and its length. Counter to our intuition, we further see that estimating mobility paths in suburban areas is more accurate. We speculate that this is due to the fact that sparser cellular network deployments occur in areas with sparser road infrastructure, thus limiting the number of possible roads that could have been taken by the connected device. In summary, we find *Cell** able to provide sufficient accuracy in path estimation for the billions of connected mobile devices that will be seen in the future.

The remainder of this work is structured as follows: we first present the problem formulation and an overview of our methodology that is composed of five steps. In Section 3, we present the details concerning each of these steps. In the following section ((§ 4)), we describe the probe that we implemented and the datasets that we collected whereas in Section 5 we evaluate the quality of the paths that are produced by *Cell**. A detailed discussion about the implications of *Cell** is given at Section 6. Finally, in Section 7 we compare with existing related work.

2. PROBLEM AND APPROACH

We begin with a definition of the problem we explore and an overview of the proposed solution.

[1]Notice that handovers are typically associated with an active communication. In our case we loosely use the term to indicate that the user switched between two sectors.

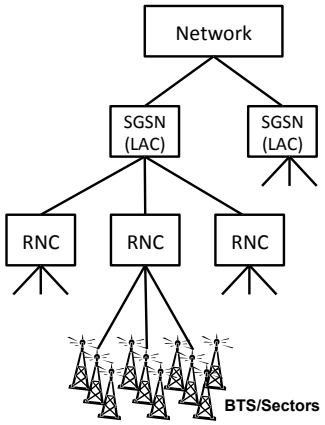


Figure 1: Overview of GSM hierarchy.

2.1 Formulation

We formulate the problem of estimating mobile paths from operational cellular data as follows. For each mobile device m or subscriber in the network, we have a sequence of time-stamped sector associations:

$$C_m = [(t_m[0], S_m[0]), (t_m[1], S_m[1]), \dots] \quad (1)$$

Each element $S_m[i]$ is the sector the mobile m is associated with at the time $t_m[i]$. A graphical representation of such a sequence of observed sectors is shown on Figure 2a. This sequence represents what can be collected from the cellular infrastructure at the BSC/RNC as part of normal operation. We note that cellular networks do not presently log data at this granularity because of its significant volume over time and the network overhead that might be imposed. Information may only be kept when a device switches across different RNCs or even LACs (Figure 1). However, it is technically and economically feasible to keep information at the BTS level if it proves useful.

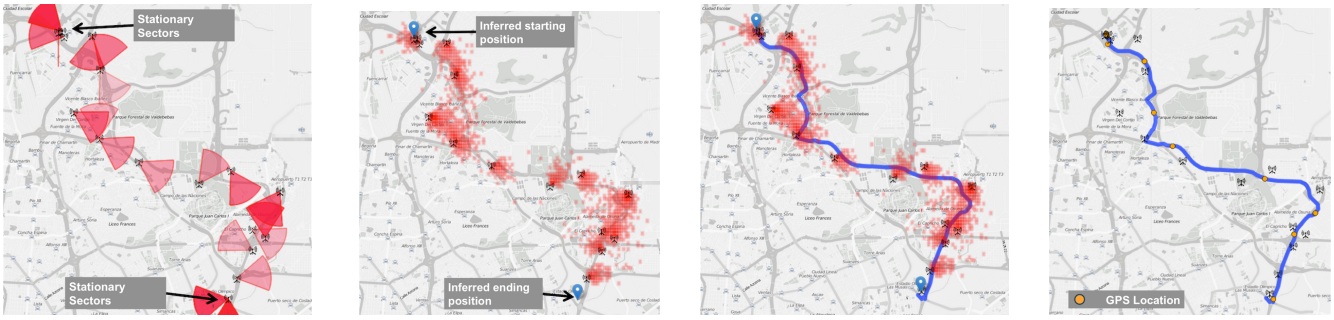
The sequence C_m is a proxy for the true location of the mobile m over time, which we call the ground truth. Ideal ground truth is continuous and accurate. However, to collect or evaluate it in practice, ground truth must be sampled at discrete times by using another localization method. Thus we model ground truth in the form we will use it in our experiments as:

$$L_m = [(t_m[0], l_m[0]), (t_m[1], l_m[1]), \dots] \quad (2)$$

where $l_m[i]$ is the location of the mobile device m at the time $t_m[i]$. Later, we will collect ground truth location using GPS as it is the most accurate sensor that is available to us.

Our goal is *to use the cellular observations C_m for each subscriber m to construct an estimate of the path of the mobile device $\hat{P}_m$ that is close to its actual path represented by L_m*. Observe that we have simplified our formulation for convenience. Both C_m and L_m are discrete trajectories that give a sequence of mobile positions at given times. We estimate the continous path $\hat{P}_m$ of the subscriber that gives its range of positions but *not the exact times* at which it progressed along the path.

This subtle change from trajectories to paths both meets our needs, since often we simply care about the street-level path, and makes it more straightforward to assess whether our estimates are close to the ground truth. We formally define what it means for estimated paths to be "close" to ground truth later on.



(a) Sequence of observed sectors. (b) Extracting start/end points and intermediate street probabilities (c) Cell* through the weighted streets (d) Comparison with GPS ground-truth

Figure 2: Cell* Steps [Best viewed in Color].

2.2 Challenges

There is much work on localization, yet the problem we explore is significantly more difficult than traditional formulations because the input data is much more limited.

Firstly, The spatial granularity of the location information is at best that of a sector/BTS (Figure 1), which is often quite large (especially in rural areas). Most significantly, there is no information on the set of cell sectors that are within range of a subscriber and there is no measurement of the signal strength (i.e., RSSI) with which a mobile can hear its associated cell sector or vice versa. While some of this information does reach the BSC/RNC when there is an active communication (i.e., data exchange), logging it would be a major departure from current operational practices such that we do not consider it a possibility in this paper. Given these restrictions, we cannot use popular localization methods that use WiFi signal fingerprinting or cell tower signal attenuation models.

Furthermore, the sequence of associations is logged at a lower granularity. Firstly, the mobile terminals may switch between BTSs of the same RNC without even notifying the network (Figure 1), making this an undetectable operation by the network. While modern smartphones frequently exchange frames with the associated towers (e.g., periodic e-mail checking), in practice, in most cellular deployments it is only viable to log information at the SGCN level: tower associations may only be recorded when a user is making/receiving calls, sending/receiving SMS, crossing Location Area Codes (LACs) or when polled by the network (usually every 2-3 hours).

Finally, even if the associated BTS is known, deriving the exact location of the mobile terminal is still challenging. Often, to balance load the BSC/RNCs may not associate mobiles to the nearest sector. Moreover, a mobile terminal itself may not necessarily switch to the nearest sector, even when it would provide a better signal, as long as it maintains a good connection with the previously associated tower. Therefore, the associated sector is partly a function of the connection history, as well as the current load, the topology and the availability of certain technologies in the area (2G/3G/4G).

As a consequence, our problem is significantly more challenging, due to the inherent sparse nature of the data under consideration.

2.3 Solution Overview

We divide our mobile path estimation procedure into five steps. We briefly explain the goal of each step and why it is appropriate. The next sections describe the methodology of each step in more detail.

- **STEP 1: Model the coverage area of the sectors.**
  To interpret mobile associations with cell sectors, we first need to model where a subscriber is likely to be while associated. That is, we need to understand the coverage of the cell sectors. For each sector, we build a model for the expected coverage based on the information that is available to the network operator. This information includes the sector location, orientation, beam width, and the intended spatial extent of the installation (i.e., macro, micro, or femto cell). A visualization of these characteristics (location, beam width, orientation and power) for an observed set of sectors C_m is shown in Figure 2a. The coverage models can be combined to identify *the overlap* between two sectors S_i and S_j to understand likely handover locations.
- **STEP 2: Identify stationary and mobile segments.**
  We model mobile subscribers as alternating between stationary and mobile states. Since different estimation procedures are appropriate for the different states, we split the cellular observations C_m into *stationary* and *mobile* sub-sequences. A stationary sequence is defined as the sequence of cell sectors S a mobile device sees while in the same logical location (e.g., building) for more than τ=15 minutes. A mobile sequence is defined as the sequence between two stationary sequences (Figure 2b).

  Separating stationary and mobile segments is not trivial because devices may typically connect to multiple towers even when they are stationary. By using the coverage maps, we can understand when handovers are consistent with a single underlying location and when handovers imply significant motion as the more likely behavior.
- **STEP 3: Estimate location for a stationary sequence.**
  To generate a path $\hat{P}_m$ we need to identify the two end-point locations. During each stationary sequence (derived from step 2), a number of sectors may be observed: the median is four unique sectors. Therefore, we can use the whole set of observed sectors over time to derive a more accurate estimate of the device's fixed location than using the whole coverage area of a single sector. Our procedure uses the coverage area of each tower (step 1) plus the duration the mobile was associated with each sector. Figure 2b shows this estimated location for the two endpoints (stationary periods) of a path. In this example, at each endpoint 3 unique sectors where used to approximate the device's stationary location.

- **STEP 4: Identify high-probability areas of the path.**
  For each mobile sequence (given from step 2), a number of intermediate sectors are observed (median is 12 unique sectors) between the stationary sequences that serve as endpoints (Figure 2a). The coverage of the observed sectors indicates the areas that the mobile device traveled through with high probability. An illustration of these inferred coverage areas of the intermediate sectors is shown in Figure 2b. We find that these areas typically constrain the path of the mobile subscriber, as shown in Figure 2c. Note that there are tradeoffs between accuracy and completeness. We find it more useful to estimate incomplete but high-probability areas such as handovers at the intersection between adjacent sectors rather than estimate complete but lower-probability areas such as the entire coverage area of all associated cell sectors.
- **STEP 5: Estimate the path with map information.**
  The previous steps give us estimates of the starting and ending locations of the path of a mobile subscriber (step 3), plus high-probability intermediate areas (step 4). We also know that humans typically move on streets or public transport routes and take fairly direct trips. Thus we use the high-probability areas to bias route search over mapping (GIS) information. More specifically, we apply the A* [16] street-routing algorithm over a map where the streets have adjusted weights based on the inferred coverage (Figure 2b). Figure 2c shows the inferred path through these high probability areas. The combination tends to work well. People typically take short paths but frequently do not take the *exact* shortest path, so a straightforward map lookup does not work well. Using the high-probability area as a bias factor captures the idiosyncratic deviations that people make from shortest paths. Figure 2d shows an example where the user did not take the shortest path.

Notice that in the above process there is no use of the timestamp $t_m[i]$ information. These can be potentially used to estimate the average speed of the device (e.g., to differentiate between highways and secondary roads) and the means of transport (e.g., walking v.s. driving). This information may be provided to the A* algorithm to improve its accuracy. However, the way that the associations are logged exhibits significant temporal noise, especially for shorter trips. More specifically, we cannot be certain about the exact time that the user departed and arrived to a certain destination: this can be only detected when there is network activity and it can happen minutes or even hours before the departure or after the arrival. Furthermore, given the large coverage area of a tower we do not know the exact location where an association took place and, therefore, calculating an average speed between events is challenging. As such, *Cell** is not making use of timestamp information.

3. SOLUTION DETAILS

We present our solution by detailing each of the five steps (§3) in turn.

3.1 Model the coverage of sectors

We build a coverage map to estimate the likelihood that a mobile subscriber m is at a particular position given the fact that it is associated with sector S. It is convenient to use a polar coordinate model for positions. The likelihood is then a function of three parameters: the power P of the sector; the distance d between the mobile and the cell tower; and the angle factor ϕ of the mobile relative to the sector. The angle factor is defined as the ratio of the polar angle of the mobile relative to the major axis of the sector and the beam-width of the sector. Thus $\phi = 0$ represents the major axis of the sector while $\phi = 1$ represents the nominal extent of the antenna beam.

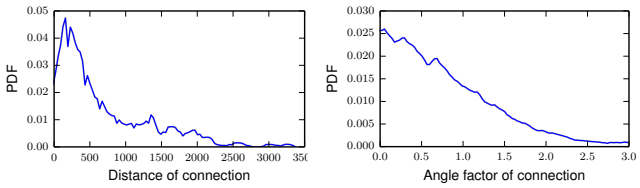


Figure 3: Projections of the coverage PDF for a macro-sectored tower on (a) distance of connection and (b) angle factor of connection.

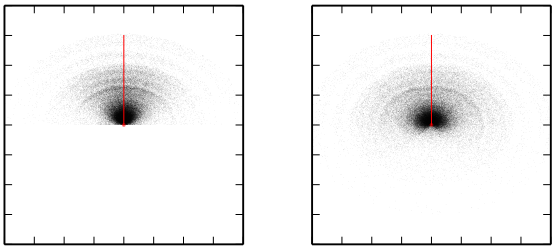


Figure 4: Coverage PDF for sectors having beam width of 60 degrees (left) and 120 degrees (right).

Rather than compute the coverage map analytically, we use ground truth data of associations at different locations to approximate it empirically. We find this to be simple and effective. (We describe how this data was collected and cleaned in §4.) The result is the coverage probabilities as a PDF that we denote by $W_P(d, \phi)$, using the same notation for parameters as explained above.

To get a sense of coverage, the projection of $W_P(d, \phi)$ on both dimensions for a macro-sectored tower is shown in Figure 3. It shows how the association probability falls off roughly exponentially with distance d (left side) and more linearly with angle factor ϕ (right side). We treat the coverage map as fixed even though the probability of associating to a particular sector changes with RF conditions, as this works sufficiently well in practice[2].

Figure 4 shows the coverage PDF directly for sectors with beamwidths of 60 and 120 degrees. For the efficient computation, we discretize space into 15m $\times$ 15m grid squares over the entire map. For every square g in the grid we derive the probability of connecting to the sector S by integrating the coverage PDF $W_P(d, \phi)$ over all points x in the square g.

$$Q(g, S) = \int_{x \in g} W_P(d_x, \phi_x)$$

We call $Q(g, S)$ the *probability grid*. In the steps that follow, we work with these squares as the most convenient form of the coverage map.

Recall that we have as input a discrete trajectory in the form of a sequence of time-stamped cell sector associations. We want to split this sequence into stationary and mobile segments. To do so, we build on the coverage PDF in the form of the probability grid $Q(g, S)$.

[2]For example, weather condition changes will affect the probability of attachment of the mobile to different sectors.

3.2 Identification of stationary and mobile segments

We call two sectors S_1 and S_2 *adjacent* if there is a square g in the grid such that $Q(g, S_1) > 0$ and $Q(g, S_2) > 0$. That is, a mobile can connect to both sectors without moving. We say a sequence of sectors $S_1, \ldots, S_n$ is *stationary* if every two sectors are adjacent and the duration of the sequence as determined by its timestamps is greater than τ. That is, the mobile can connect to all sectors without moving and this situation persists for long enough that non-adjacent handovers would likely have occured if it was actually moving.

Finally, we extend stationary periods as far along the sequence as possible. Then we define the sequence of sectors between two stationary periods (which cannot itself be stationary) to be a *mobile* period. In our data (see §4), we find that many subscribers connect to a single sector for long periods of time in a stationary period. The median number of connected sectors in a stationary period is 4. In contrast, in mobile periods the subscriber tends to associate to multiple non-adjacent sectors with frequent handovers. The median is 12 sectors per mobile period.

3.3 Finding the location while stationary

We use all information in the stationary period to estimate the stationary location because an accurate endpoint is instrumental in inferring an accurate route for the mobile segment.

If a stationary segment consists of only one sector S, the most likely position is simply the square g that maximizes $Q(g, S)$. There is little we can do to improve the accuracy of the estimate. However, when the stationary period contains several sectors $S_1, \ldots, S_n$ then we can use all of the sectors to refine our location estimate. To do so, we find the point g that maximizes the sum of probabilities $Q(g, S_i)$ over the sectors weighted by the amount of time τ_i the subscriber was connected to the sector S_i. That is:

$$g^* = \arg\max_g \sum_{i=1}^{n} Q(g, S_i)\tau_i.$$

For a complete path consisting of a starting stationary period, mobile period, and ending stationary period, this procedure gives us an estimated location for the start and end points. We note that the end point for one path forms the start period for the next path with no discontinuities because they share the same stationary period.

3.4 Likelihood of crossing a square in the grid

While we have an estimate for the stationary locations, we have not yet used the information in the mobile period. As a final input to our path estimation procedure, we mine the mobile sequence for locations that the subscriber passes through with high-probability. In the next step, we will bias our estimated path to follow the high-probability locations.

Recall that for a square g, $Q(g, S)$ is the probability that a subscriber in g is connected to S. For each square g we denote the *score* $\rho(g)$ of the mobile segment $S_1, \ldots, S_n$ as:

$$\rho(g) = \sum_{i=1}^{n} Q(g, S_i). \quad (3)$$

That is, we compute the union of the coverage PDF for all sectors in the mobile segment. This score is no longer a probability, but the higher the score for a square g the more likely that the path of the subscriber passed through that square. Observe that summing the coverage PDFs also increases the weight on locations that overlap between sectors by "double counting" them. These are precisely the most likely handover locations.

3.5 From mobile segment to a path

Our early efforts at path estimation showed us that purely geometric approaches based on cell tower locations were unlikely to yield good accuracy. (We evaluate this in §5.) The location problem is simply too under-constrained. To improve accuracy, we observe that real paths are consistent with the underlying road infrastructure. We can use the road network to constrain the path between the stationary endpoints.

An obvious question is whether, given the use of GIS maps, we need to use the mobile segment at all – we do not if people simply follow the shortest (or fastest) path on the road network between two locations. Zhu and Levinson [23] demonstrate that humans do tend to follow short paths when traveling between two locations, but with variations on the shortest route owing to many attributes such as how they value their time, willingness to pay tolls and fuel, time budgets, behavioral preferences, habituation, and experience of a given route. They show that 60% of the users take the exact shortest path as their route, and in almost 90% of cases people choose routes that have small deviations, of at most 5 minutes away from the optimal one.

Our approach is to use the scores of the squares $\rho(g)$ to bias routing on the road network. By using the scores, we preferentially search for a route in the "corridor" formed by the coverage areas of all sectors in the mobile segment. High scores close to the shortest path represent likely deviations.

Our implementation uses OpenStreetMaps (OSM) [10], a community-driven repository that contains data about roads, POIs, railway connections, etc. OSM is designed to be used for navigation purposes as it includes information for routing by many modes including car, foot and bicycle. The OSM data can be thought of as a graph $G = (\mathcal{V}, E)$ of the road network. Nodes $\mathcal{V}$ represent the geographical locations of intersections or points along curved road segments, while each edge represents a directional road-segment between two nodes. In the rest of the paper, we use the car OSM mode, and leave the problem of also inferring the mode of transport for future work. We proceed as follows.

Initial edge weights: For each edge $e \in E$ an initial weight, $\mathcal{W}(e) \in \mathbb{R}$ is assigned that represents the *expected time* that is required to traverse it. This weight depends on: i) the length of the road segment; ii) the type of the road segment (e.g., motorway, primary, secondary, footpath); and iii) the mode of transport (e.g., walking, driving). The required information and the default values of these weights for each mode of transport are calculated based on the OSM recommendations [10]. Notice that for some means of transport certain roads might be restricted (e.g., motorways cannot be used for walking routes).

Modifying the weights based on mobile segment sectors $S_1, \ldots, S_m$: After calculating the default weights for every road segment, we adjust them based on the scores of squares in the grid. Each road segment e crosses several square grids $g_1, \ldots, g_l$. We want to up-weigh roads with high scores and down-weigh roads with low scores to prefer locations that the subscriber crossed with high probability. We adjust the weight of the segment e heuristically as:

$$\bar{W}(e) = \frac{W(e)}{\frac{1}{l}\sum_{i=1}^{l} \rho(g_i) + \epsilon \max \rho(g)}$$

The use of ϵ allows the route to gracefully handle gaps in coverage due to low (zero) scores rather than require a major diversion. It basically allows to balance the tradeoff between the short/best routes and the routes that are close to the observed sectors. We experiment with various ϵ values and observe that as long as $\epsilon \leq 0.1$ the expected errors reported in the evaluation section are insensitive of the choice of ϵ.

Routing: After adjusting the weights, we use the standard A* algorithm to search the graph for a shortest weighted path between the start and end points. We use A* for pathfinding because running Dijkstra on the entire network graph is computationally expensive; A* is much faster and performs well in our setting. The resulting path in the graph is our final estimate for the path taken by the mobile subscriber, i.e $\hat{P}_m$. An example is shown in Figure 2d.

4. EXPERIMENTAL DATASET

To evaluate *Cell**, we need a dataset that contains cellular-side observations and the corresponding ground truth for the paths taken by mobile subscribers. While our algorithm is designed to work solely with cellular-side data, we find it expedient to collect matching trajectories for cellular data and location ground truth using a smartphone. We describe the collection method and dataset in this section.

4.1 Probe Application

We implemented a probe application for Android and used mobile phones to collect a dataset for evaluation. The probe runs as a background service collecting either event-driven or periodic measurements. Measured samples are stored locally at the device as a collection of geographically-annotated, time-stamped GeoJSON objects. They are periodically uploaded to our servers. The most important types of measurements are summarized below.

Location: A location sample is generated whenever there is any change to the *network-based* (WiFi or GSM trilateration) location of the device as reported by the Android API. In addition, we *periodically collect GPS samples*. To avoid depleting the user's battery, *one GPS sample per minute* is collected and the GPS sampling is paused while the user is connected to a WiFi network and associated to the same cell (e.g., when stationary at work/home).

Cellular Information: A sample is generated every time there is a handover from one sector to another. The collected information includes the IDs of the sector, RNC, and LAC (see §2). This information can be used to identify the tower in the database of the cellular operator. Notice that additional information, such as periodic collection of the full lists of received towers and the RSSI to each one of them, is also collected by the probe, but it is not used for this study because this information cannot readily be logged by the cellular operator. Finally, any connectivity changes or switches between types of connectivity (2G/3G/4G) are also logged.

Wi-Fi Information: Events concerning the Wi-Fi connectivity changes (e.g., connections and disconnections) are logged. This information can help us to further geo-locate the mobile in post-processing when the accuracy reported by Android is not adequate (e.g., due to a lack of connectivity to query Google servers) and to understand if the user was in the same indoor location.

4.2 Dataset and Cleaning

The probe was installed by 30 users and used over a period of 8 months[3]. All users live within a single country but multiple cities are covered. To increase diversity we selected users that cover a variety of age groups (20-60), who live in both urban and suburban areas, and who exhibit different commute patterns (bus, walking, cycling, driving, underground). In total, we logged more than 4.6 million samples over 19,000 hours of collection. A break-down of these samples is shown in Table 1.

Type	Number
Total number of Cell samples	1,718,504
Total number of Hand-overs	433,031
Number of Distinct Sectors observed	15,455
Sectors in the operator's database	$>$ 100,000
Number of Location samples (Netw. and GPS)	673,468
Number of GPS samples	259,032
Total time logged (all devices)	19,438 hours
Total time stationary	18,335 hours
Total time moving	1,102 hours
Number of trajectories (trips)	3,216
Total distance traveled	19,840 km

Table 1: The collected dataset. Each sample is a record of a sector association, handover, GPS/network location update, etc.

Before using the dataset for experiments, we had to clean it to remove outliers. We proceeded as follows.

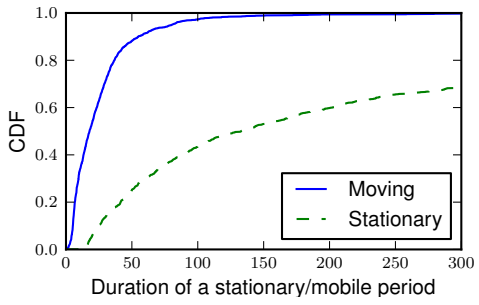
Inaccurate GPS: GPS or network-based location is used as the ground truth in our study but it does have its own accuracy errors. Android reports location information along with an estimate of its error L_{error}. In our samples, we observe a 90th-percentile error of 86 meters for GPS (median is 19 meters). This compares to 142 meters for indoor network-based locations (median is 65 meters). To limit problems, we discard any location measurements with reported error $L_{error} \geq 100$ meters. These inaccurate samples make up 8% of the 673,468 location samples.

Stale sector database: We have access to a database for all the sectors of a major cellular operator that contains the exact location, orientation, beam-width, power characteristics and type of installation of more than 100,000 sectors within a country. The database is updated continually. For instance, we notice that more than 100 sectors are re-located monthly whereas more than 3000 cells are added or removed. To protect against stale data, we remove any associations to sectors that have moved or are obviously misplaced (e.g., more than 20km away from the GPS location of the user). This cleaning process removed only 0.06% of the 433,031 observed associations.

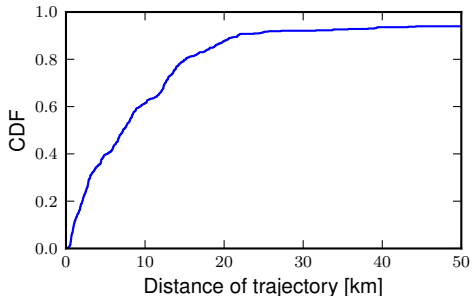
Gaps in collection: Gaps in data collection may happen occasionally because smartphones might run out of battery or terminate the probe application. The difficulty with gaps is that they can lead to unrealistic trajectories without a normal start or end. We mark these gaps so that we can filter out any trajectory that includes them.

Subway usage: In the country that we collected the data, metro (underground) stations have cellular coverage through a *radiating cable* that runs along the whole metro line. This means that while a user is traveling within a single metro line, she is always associated to the same sector. Thus different methods are required for estimating trajectories that include metro usage. Since the metro is not our focus, we identify and remove any sectors that cover metro lines.

[3]Some users did not run the probe for the full duration of the experiment.



(a) CDF of interval duration.



(b) Distance of a trajectory.

Figure 5: Ground-truth statistic.

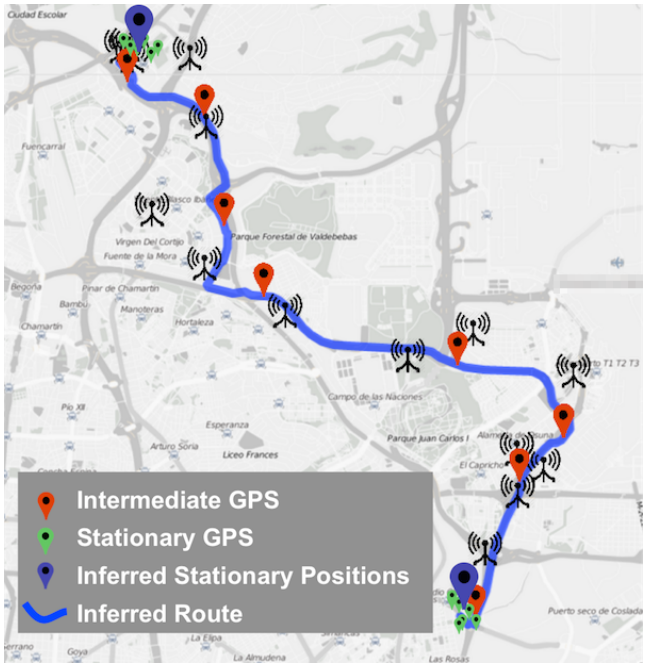


Figure 6: GPS ground-truth locations and estimated path

4.3 Ground-Truth Trajectories

We need ground-truth trajectories to evaluate how well *Cell** estimates the paths. Notice that even when a user is stationary, the reported GPS or network (WiFi-based) locations can exhibit small variations over time. An example is given in Figure 6. Therefore, to identify when the user is stationary and to find the actual fixed location we follow the *Leader-based Clustering* methodology described in [14]. In essence we merge nearby location samples: for every location sample, we determine if it belongs to any of the already generated clusters by computing its distance to the cluster leader's location. If this radius is below a threshold $R_{cluster}$ the point is added to the cluster and the most central point (median) is the new leader. At the end of this process each cluster contains all the location samples that are within $R_{cluster}$ of the center of the cluster.

To determine the threshold (the cluster radius) we used the WiFi connectivity and the distribution of L_{error}. We considered a user stationary when they stayed connected to a specific WiFi BSSID (e.g., at home or work). We determined that $R_{cluster} = 100m$ produced clusters that correctly contain such intervals. This value is also consistent with the L_{error} that we used for GPS data-cleaning.

To find trajectories, we consider the user in a stationary period, that represents an endpoint, if the time that the user stayed in the cluster is larger than τ. For consistency, we use the same $\tau = 15$ minutes as with *Cell**. Users are in mobile periods between stationary periods. This procedure identifies 3,216 distinct trajectories from our dataset.

Figure 5a shows the distribution of the time that a user was in stationary and mobile states using the ground-truth. Mobile trajectories are much shorter in duration than stationary periods. The median duration for a mobility segment is 23 minutes, and the majority of all recorded trajectories do not exceed one hour. In Figure 5b we observe that most mobile trajectories are shorter than 10 kilometers. On the other hand, stationary periods have a median duration of 132 minutes, while 22% of them last for more than 8 hours. Such a result is expected because most stationary segments correspond to the user being at work during the day or at home during the night.

5. EVALUATION

In this section, we use our experimental dataset to evaluate the quality of the paths that are produced by *Cell**.

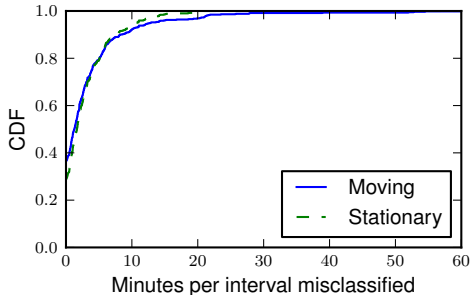
5.1 Methodology

We want to know how closely the paths $\hat{P}_m$, output by *Cell**, match the true paths of mobile subscribers. Recall our situation as shown for an example trajectory in Figure 6.

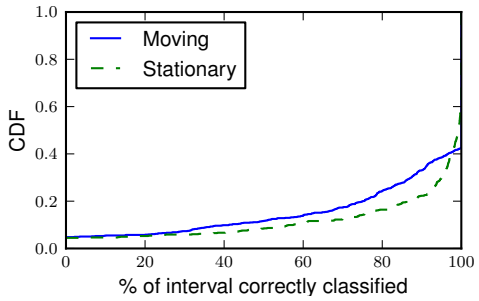
The subscriber follows a true path that is not shown in the figure or known to us. What is known is the measured GPS trajectory L_m. The GPS samples of this trajectory are shown in green (near the ends) for stationary periods and red (in the middle) for mobile periods. We use L_m as ground truth for our evaluation since it is the best estimate of the true path that we can gather.

However, note that the GPS samples are not true locations. They have an accuracy of L_{error} meters that is collected as part of GPS sampling. Even if the user is stationary the recorded location may slightly vary over time (especially when the user is indoors). The median reported accuracy is 19 m. This implies that, even if *Cell** estimates the mobile path perfectly, a comparison with GPS-based ground truth will show a residual inaccuracy.

The cellular trajectory C_m is denoted in the figure by the cell towers used for each handover. *Cell** uses these observations to produce the estimated path of the subscriber $\hat{P}_m$ that is shown as the solid blue line. While there are many ways that we might measure the accuracy of estimated paths, we use a simple method that we found to be sufficient. For each GPS sample, we can compute the shortest distance between the GPS location and the estimated path. We denote this error sample e_i. It represents how close the

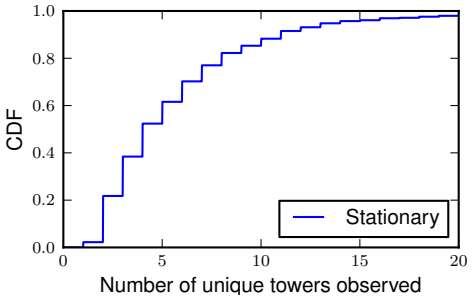


(a) CDF of time that was misclassified per period.

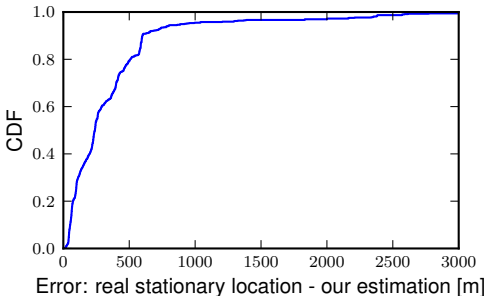


(b) Percentage of a period that was correctly classified.

Figure 7: Accuracy of detecting stationary/mobile periods.



(a) # of unique sectors observed.



(b) Error of estimated stationary location.

Figure 8: Accuracy of estimating stationary location.

path goes to the GPS location, and we look at the distribution of error samples across all estimated paths. We also compute the error samples for a complete estimated path so that we can look at accuracy at the level of individual paths. We denote this sequence of error samples for an estimated path E_m.

In the next subsections, we evaluate the steps of our procedure. We begin by seeing how well *Cell** identifies the stationary and mobile segments of the trajectories (§3.2). Then we compute the accuracy of the stationary locations compared to GPS locations (§ 3.3). Finally, we look at the accuracy of the estimated paths (§5.5).

To help understand why the steps in our procedure lead to high accuracy, we contrast *Cell** with two methods that estimate paths with less information: a method that incorporates the most likely location of the mobile while associated with the recorded sectors, but that does not use any GIS information (just connects the points); and a pure map method that finds a route between the endpoints while ignoring all intermediate sector information. These comparisons indicate that a combination of methods is needed. Finally, we characterize the conditions under which *Cell** operates well.

5.2 Accuracy of static and mobile segments

In §3.2, we separate the sequence of sector observations into stationary and mobile periods to form trajectories. We use the ground-truth GPS trajectories L_m to evaluate the accuracy of this process. We use three measures: i) the time that is misclassified; ii) the percentage of a ground-truth stationary/mobile period that is misclassified; and iii) the number of trajectories that could not be identified.

Figure 7a shows the distribution of time (in minutes) that is misclassified per ground-truth stationary/mobile period. The median error is 2 minutes and the 80th-percentile of errors is less than 5 minutes. These results are quite good given that we are limited by coarse-grained cellular-side data that only captures user movement when a handover occurs. As we will see later, most of these errors occur during the beginning and the end of a mobile period when the user is still connected to the same sectors as in the stationary periods.

In Figure 7b we plot the distribution of the percentage of the ground-truth period that was correctly identified. We observe that we accurately identify a large percentage of the stationary and mobile periods. For each of 80% of static periods and 70% of mobile ones, we achieve more than 90% of accuracy.

Finally, only 3.1% of the mobile trajectories are not identified at all by using cellular-side data. Further investigation showed that these are short trips within the same coverage area; we cannot identify trajectories without handovers.

5.3 Accuracy of stationary locations

Next, we evaluate the accuracy with which we estimate the stationary location of a subscriber. These locations are important because they form the endpoints of the estimated path. In §3.3, we describe how we exploit multiple sector observations to narrow down our estimate of the stationary position of the subscriber. Figure 8 shows the distribution of number of unique towers observed during a stationary period. In most of the cases (98%) a user is associated to 2 or more sectors while the median is 4 sectors. This confirms that we do stand to improve accuracy by combining multiple cellular observations for the stationary location.

We define the error in stationary location as the geodesic distance between the *Cell** estimated location and the median of GPS locations during a stationary period. The CDF of errors is plotted in Figure 8b. About 50% of estimated locations have an error smaller than $230m$, and 80% of the estimated locations have an error less than $500m$. These results are significantly more accurate than the median distance of $480m$ at which a mobile subscriber is associated to a tower.

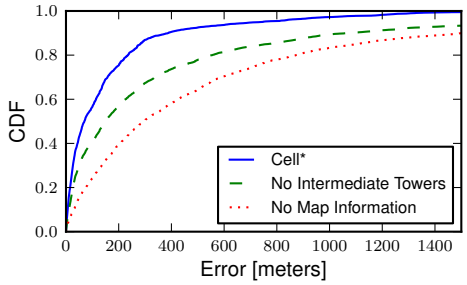


Figure 9: Error between GPS locations and estimated paths.

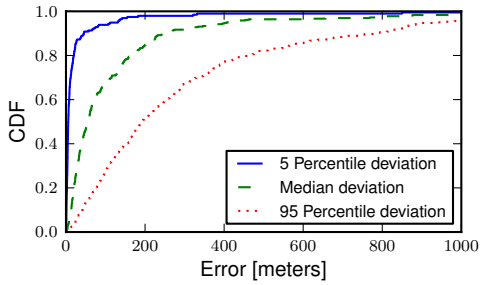
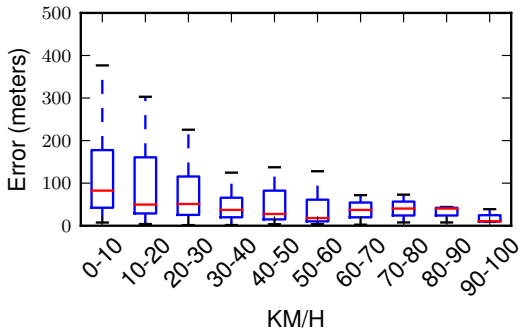
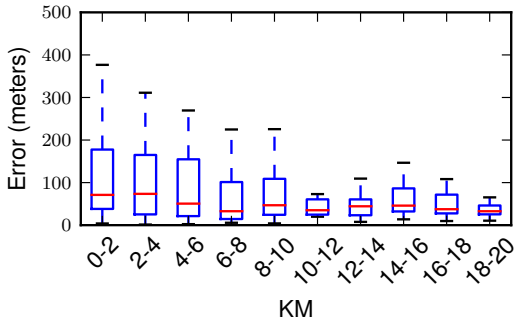


Figure 10: Per-path error statistics.



(a) Path median error vs its Speed.



(b) Path median error vs its Distance.

Figure 11: Median error per route.

5.4 Accuracy of estimated paths

Now we evaluate the quality of complete paths produced by *Cell**. To get an overall sense of accuracy, we compute the error samples e_i that are the shortest geodesic distance from each GPS location associated with a trajectory to the estimated path. Figure 9 shows the distribution of the error samples across all trajectories. We observe that 75% of the errors fall below $180m$, and the median error is only $70m$. This positive result shows that *Cell** can accurately estimate paths given only the set of cellular handovers and without the assistance of GPS. For comparison, the median GPS accuracy for the same data is $19m$.

To understand how the different steps of our method contribute to the accuracy of paths, we make two comparisons. First, we run the street-routing algorithm on the road network between the stationary endpoints with basic weights of the roads according to their types. This process is equivalent to asking directions on Google maps or any other navigation system; no intermediate points due to handovers are considered. As we observe in Figure 9, while the median error is doubled to $150m$, many of the errors are reasonably low. This implies that street-routing provides valuable semantic information for finding real paths. However, 10% of the errors have large deviations of more than 1km. Street-routing is not sufficient by itself – it can go very wrong which is why it is important to use intermediate handovers.

Second, to see the value of using the intermediate handovers alone, we compare *Cell** paths to a path that simply connects the observed trajectory sectors with a straight line. The errors for this simple path are also shown in Figure 9. We see that they are significantly larger than *Cell** or even street-routing. The median error is 298 meters and in 10% of the cases we find deviations of more than 2.5km. We conclude that street-routing is essential.

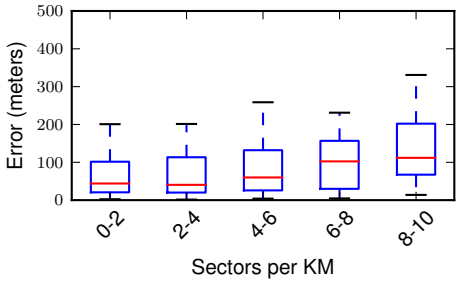
We can also look at error samples on a per path basis. For each estimated path, we compute the 5th, median, and 95th percentile of the sequence of errors for the path E_m. In Figure 10 we see that 50% of the estimated paths have a median error of just $54m$, and a 95th percentile error of $200m$. If we consider that the average block size of a city is approximately 200 meters, we are able to estimate paths plus or minus one block 95% of the time.

5.5 Accuracy by Type of Path

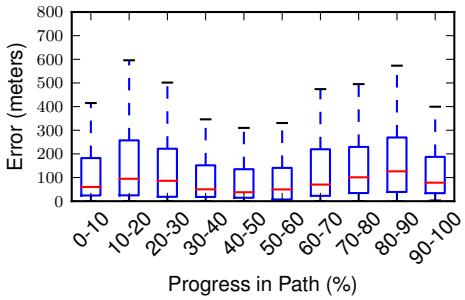
Finally, we look into path properties that may make their estimation easier or harder. One such property could be the speed. High speed paths are likely to generate a larger number of handovers, and be constrained in a smaller area of the road network like arterial roads. Another related factor is distance, since longer paths generate a larger number of handovers.

Figure 11a shows the distribution of median errors per path when the paths are clustered according to their average speed. We observe it is harder to infer low-speed paths, as these are typically more chaotic walking routes in dense, urban environments. (They may also require the use of a different mode of OSM). For instance, some of these routes do not have specific destinations as they are random city walks, searching to shop etc. In contrast, we can accurately estimate high-speed paths as these typically include highway segments that are easier to predict.

Much the same behavior is observed when we look into how distance affects the error (Figure 11b). We are less accurate to estimate short paths as these include very few sector handovers, and the selection of route may deviate even farther from the absolute shortest path.



(a) Path median error vs Density of deployment.



(b) Path median error vs Stage within the route.

Figure 12: Median error per path (meters).

It is also evident to us that the errors in path estimation are impacted by the density of the cellular deployment in the vicinity of travel. For example, suburban areas are expected to be covered with far fewer sectors than densely populated urban areas, and to cover a far greater distance with one sector. Indeed, the highest errors in our static location estimation come from suburban areas. To see density effects for the whole path, we define sector density for a path as the number of sectors available throughout the path divided by the path length.

Figure 12a shows the relationship between the path median error and its sector density. Contrary to our original intuition, we have more accurate path estimates in less dense tower installations (e.g., suburban areas with less than 0-2 towers per km). We conjecture that this happens because suburban areas tend to have sparser road networks as well, and their inhabitants tend to use cars more than in urban areas. We leave an exploration of the correlation between road density and cell infrastructure density for future work.

Finally, in Figure 12b we plot the accuracy of the estimated path as a function of the normalized elapsed time from the beginning of the path. Interestingly, *Cell** is more accurate during the initial and final 10% of the route. This happens due to the fact that we have a good estimation for the start and ending points. Furthermore, the accuracy is also good for the middle part of the trip as most people use primary roads (e.g., highways). It is slightly worse transitioning between these situations, likely due to less predictable secondary roads.

6. DISCUSSION

*Cell** enables mobility path estimation for connected devices that feature no GPS nor WiFi due to size, cost, or energy constraints. It can further be used on anomymized cellular network data in order to get information about the mobility patterns of the millions of subscribed devices, that could span both the "things" of the future,

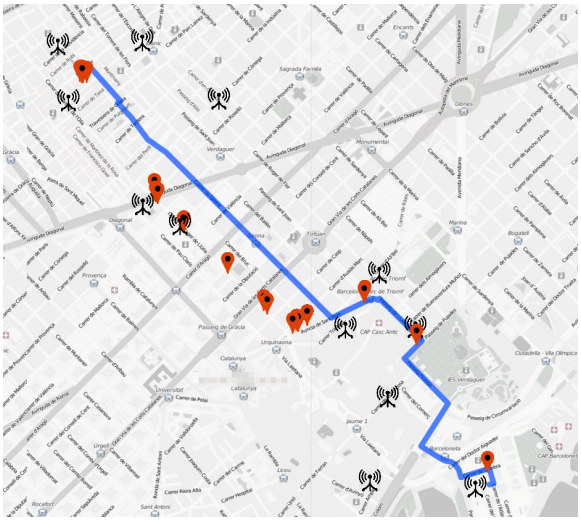


Figure 13: Path estimation can sometimes be one block off from actual taken path

but also smartphone users. One possible implementation could be through a network API, that given the IMEI could return the estimated path for a predefined period of time.

Our results demonstrate a median accuracy of less than 70 meters, that falls within an average size city block, i.e. in 50% of the cases, our path estimation may take the road parallel to the one actually taken by the device. Figure 13 shows an example of such a case in our evaluation data - a part of the estimated path deviates from the actual path by one single block. As a result, *Cell** could enable a number of novel services for mobile connected devices, but may not be able to estimate the exact path taken under all conditions.

We find that the applications of mobility paths with such a level of accuracy could still be numerous. Examples could span city transportation planning, traffic monitoring at the macro level, or fleet management tracking in a computation and energy free fashion through an opt-in service, where the cellular network acts as a repository of mobility information.

Despite the fact that mobility path accuracy is not always to the exact street level, we find that higher order metrics computed on the extracted trajectories could still allow different operations with sufficient accuracy. For instance, we have computed the total length of the extracted mobility paths and compared it with the ground truth, as computed from GPS data. We find that despite the slight deviations in the derived paths, the total distance is just within 4.1% of the actual path on average. One could envision services where the estimation of the total distance may be more important than the precise path taken (for instance within a car insurance context, where the insurance company assigns risk based on the total distance driven).

Another point of discussion has to do with the granularity of data considered. While collection of handover events from all mobile subscribers is technically feasible today, in practice it comes at a cost with respect to the required network infrastructure (although with the adoption of LTE, logging information at high granularity might become economically viable). Our data captures device GPS location, along with BTS/RNC/LAC cellular information. In what follows we quantify the loss in accuracy in *Cell** if it is provided with even sparser data, that may be accessible at the RNC level or at the LAC (Figure 1). To do that we sub-sample the location infor-

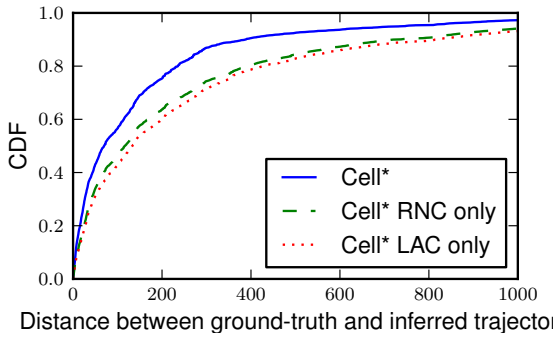


Figure 14: Path accuracy as a function of the cellular layer where location information is collected.

mation collected by our Android probe to only include handovers that are visible at the i) RNC or ii) LAC level [4]. We find that sparser observations for the mobile device lead to increased inaccuracy of 129 meters in the median case (up from 69.7 meters), while the 75th percentile error is now 340 meters (up from 200 meters). The entire distribution of errors for cellular location information collected at the different layers of the cellular infrastructure is shown in Figure 14. We were actually quite surprised that *Cell** still behaves well under such loss of information. We intend to look deeper into this in future work.

Finally, we note that the results presented in this paper capture the properties of the studied cellular provider network and the road network of the studied city. As mentioned previously, we speculate that geographic areas with dense road networks are very likely candidates for denser cellular network deployments. The tendency in the telecommunications area is that of reducing the radius of cell towers, going to micro- or femto-cell deployments. A decrease in the coverage area of cell towers and sectors will only improve the accuracy of *Cell**.

7. RELATED WORK

Along with the advance of tools to monitor fine-grained human mobility through mobile devices, our understanding on human mobility has been expanded in recent years. Researchers have found not only a basic law of governing human mobility [7] but also a limitation of predictability [21] based on the unprecedented volume of mobility traces. One of the important discoveries is a high degree of spatial and temporal regularities of human mobility [7]. It offers a strong foundation for a wide range of research challenges, such as important places identification [11], next location prediction [5] and frequent trajectory extraction [6].

Research on human mobility becomes important for urban planning, land use identification, population movement monitoring, traffic forecasting, transportation system management, travel demand prediction, and even infection control. Among many data sources for those studies, Call Detail Records (CDRs) play a prominent role due to their scalability. CDRs inherently have low spatial and temporal resolution, but a tremendous advantage - that of observing the mobility patterns of millions of cellular subscribers with no support required on the cellular device. Given the sparsity of data, most approaches working on route or location mining on CDRs are based on probabilistic approaches. Görnerup proposes a scalable probabilistic method based on locality-sensitive hashing and graph clustering for inferring common routes from sequences of sectors [8]. Saravanan et al. propose to aggregate CDRs to find people's daily routes by constructing a Gaussian model that explains the probability of people being around specific towers [19]. Also, Isaacman et al. present how to model people's movements in metropolitan-scale areas using spatial and temporal probability distributions extracted from CDRs [12]. All such approaches do not consider the geographic information but take the resolution of cell towers. By contrast, *Cell** enables to estimate the street-level path by incorporating the geographic information.

Building origin-destination (OD) matrices, which usually estimate the number of vehicles commuting between residence and workplace, is one of the popular applications [11]. In contrast to conventional methods to acquire OD matrices mainly depending on household surveys or road monitoring, using a GSM network is much more cost-effective [3, 4]. It shows the potential of huge volume of CDRs for urban planning and transportation engineering. Beyond OD matrices, that reveal endpoints of trajectories, there have been a few studies to infer a street-level path by using cellular-side information. Becker et al. use cellular hand-off patterns to identify commuting routes [1]. They generate a collection of hand-off patterns, as observed through fingerprinting by previously test driving the target area. The proposed method achieves high accuracy but the dictionary of hand-off patterns for every route is essential in advance, limiting the areas where such an approach can help.

Looking at client-side solutions, Paek et al. propose a lightweight positioning system using a mapping between GPS locations and associated cells for regular and relatively long routes [17]. In their system, each device records GPS locations and associated towers regularly. Then, a sequence of passed cells and the elapsed time can refine the current locations based on history. As this system essentially requires GPS readings as input, it does not directly compare to *Cell**.

Schlaich et al. propose a method to generate all the possible routes by using a multi-path route generation algorithm based on cellular-side location-area update messages [20]. They propose heuristics, such as taking road type into account, to restrict the selection of routes, but they do not have well defined criteria to choose the most probable route among many candidates. *Cell** neither collects specific data for each road, historical data, nor proposes all possible options. It suggests the most probable path by combining sector characteristics and GIS information.

With the near ubiquity of GPS-enabled devices, map matching becomes a core component of a wide range of applications, such as car navigation, route prediction, and activity recognition. Quddus et al. [18] conduct a comprehensive review of the map matching algorithms, from simple point-to-point mapping [2], to topological analysis of the spatial map [9], probabilistic approaches dealing with error regions [15], and sophisticated fuzzy modeling ones [13]. Cellular-side data is coarse-grained at the resolution of sectors. Thus, there are clear analogies between inferring a trajectory based on such low-resolution observations and identifying the most relevant road segment based on inaccurate GPS positioning data, although errors in distance can be quite different. *Cell** uses the probabilistic approach to refine locations and paths.

[4]Given space constraints we omit the details for cellular network infrastructure.

8. SUMMARY

Through their normal operation, cellular networks are a repository of continuous location information from its subscribed devices. Such information, however, comes at a coarse granularity both in terms of space, as well as time. For otherwise inactive devices location information can be obtained at the granularity of a BTS, and at infrequent points in time, that are sensitive to the structure of the network itself, and the level of mobility of the device. *Cell** enables the extraction of mobility paths from sparse spatio-temporal cellular location information. Using more than 3,000 mobility trajectories, we show that we are able to estimate the stationary locations with a median accuracy of 230 meters. When the device is mobile, we can estimate its mobility path with a median error of 70 meters. To achieve such accuracy, we are taking into account the cellular network topology, as well as geographic information. We show that mobility path accuracy improves with its length and speed, and counter to our intuition, accuracy appears to improve in suburban areas. *Cell** is the first technology, we are aware of, that allows location services for the new generation of connected mobile devices, that may feature no GPS, due to cost, size, or battery constraints.

Acknowledgments:
The research leading to these results has received funding from the European Union under the FP7 Grant Agreement n. 318627 (Integrated Project "mPlane"). We would also like to thank Yan Grunenburger, Enrique Garcia Illera, Hosung Park and Andreu Urrela Planas for their support and comments on this work.

9. REFERENCES

[1] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, 2011.

[2] D. Bernstein and A. Kornhauser. An introduction to map matching for personal navigation assistants. 1998.

[3] N. Caceres, J. Wideberg, and F. Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, 2007.

[4] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011.

[5] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[6] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339. ACM, 2007.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[8] O. Görnerup. Scalable mining of common routes in mobile communication network traffic data. In *Proceedings of the 10th international conference on Pervasive Computing*, Pervasive'12, pages 99–106, 2012.

[9] J. S. Greenfeld. Matching gps observations to locations on a digital map. In *National Research Council (US). Transportation Research Board. Meeting (81st: 2002: Washington, DC). Preprint CD-ROM*, 2002.

[10] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

[11] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in peopleâĂŹs lives from cellular network data. In *Pervasive Computing*, pages 133–151. Springer, 2011.

[12] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12, pages 239–252, 2012.

[13] S. Kim and J.-H. Kim. Adaptive fuzzy-network-based c-measure map-matching algorithm for car navigation system. *Industrial Electronics, IEEE Transactions on*, 48(2):432–441, 2001.

[14] A. Kirmse, T. Udeshi, P. Bellver, and J. Shuma. Extracting patterns from location history. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 397–400, New York, NY, USA, 2011. ACM.

[15] W. Y. Ochieng, M. Quddus, and R. B. Noland. Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, 2(55), 2004.

[16] N. J. N. P. E. Hart and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems, Science, and Cybernetics*, SSC-4(2):100–107, 1968.

[17] J. Paek, K.-H. Kim, J. P. Singh, and R. Govindan. Energy-efficient positioning for smartphones using cell-id sequence matching. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 293–306. ACM, 2011.

[18] M. A. Quddus, W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

[19] M. Saravanan, S. Pravinth, and P. Holla. Route detection and mobility based clustering. In *Proceedings of IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application (IMSAA)*, pages 1–7, 2011.

[20] J. Schlaich, T. Otterstätter, and M. Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies*, 2010.

[21] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[22] A. Varshavsky et al. Are gsm phones the solution for localization? In *Mobile Computing Systems and Applications, 2006. WMCSA'06. Proceedings. 7th IEEE Workshop on*, pages 34–42. IEEE, 2005.

[23] S. Zhu and D. Levinson. Do people use the shortest path? an empirical test of wardropâĂŹs first principle. In *91th annual meeting of the Transportation Research Board, Washington*, volume 8, 2010.