Can You Find Me Now? Evaluation of Network-based Localization in a 4G LTE Network

Robert Margolies, Richard Becker, Simon Byers, Supratim Deb, Rittwik Jana, Simon Urbanek, Chris Volinsky
AT&T Labs Research, Bedminster, NJ
{robertm, rab, byers, supratim, rjana, urbanek, volinsky}@research.att.com

***Abstract*—User location is of critical importance to cellular network operators. It is often used for network capacity planning and to aid in the analysis of service and network diagnostics. However, existing localization techniques rely on user-provided information (e.g., Angle-of-Arrival), which are not available to the operator, and often require a significant effort to collect training data. Our main contribution is the design and evaluation of the Network-Based Localization (NBL) System for localizing a user in a 4G LTE network.**

The NBL System consists of 2 stages. In an offline stage, we develop RF coverage maps based on a *large-scale crowd-sourced* channel measurement campaign. Then, in an online stage, we present a localization algorithm to *quickly* match RF measurements (which are already collected as part of normal network operation) to coverage map locations. The system is more practical than related works, as it does not make any assumptions about user mobility, nor does it require expensive manual training measurements. Despite the realistic assumptions, our extensive evaluations in a national 4G LTE network show that the NBL System achieves a localization accuracy which is comparable to related works (i.e., a median accuracy of 5% of the cell's coverage region).

***Index Terms*—Localization, wireless networks, crowd-sourcing.**

I. Introduction

The size and complexity of cellular networks continue to grow, with upcoming 5G networks expected to handle 1000-fold increases in the amount of traffic and 100-fold increases in the number of users [1], [2]. To address these ever-growing requirements, cellular networks are becoming ever-more complex, with operators deploying heterogeneous cells composed of macro cells, small cells, distributed antenna systems, etc. Each deployment must be controlled through thousands of configuration parameters. Thus, managing and monitoring the network through manual processes is not feasible. Instead, operators are moving towards an automated measurement-driven approach to network control and management (e.g., [3]).

As part of the network control and management, cellular operators collect extensive amounts of log information from all users and cell towers. The terminology for this data varies by vendor and operator; we will refer to this data as User Measurement Data (UMD) [4]. As shown in the left side of Fig. 1, the data is collected through the cellular network and streamed to a number of network edge locations. For example, over 3.5TB of 4G LTE UMD is collected per day in the network edge location studied in this work (owned by a top-3 network operator). The data contains a comprehensive set of measurements for every call and data session by every user. This includes metrics such as throughput, latency, and RF measurements (RSRP, RSSI, etc.) as well as diagnostics such as dropped call information. As such, UMD is a prime source of information to drive automated network control.

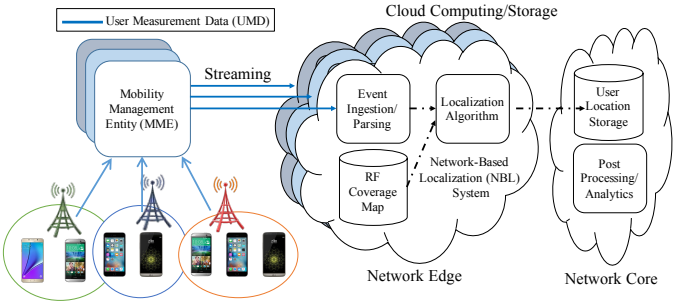


Fig. 1. User Measurement Data (UMD) collection and Network-Based Localization (NBL) System architecture.

However, since UMD is collected by the network, it does not contain any user-provided location information (i.e., GPS). Yet the value of such location information is clear; for the network operator, it can be used to (i) identify location hotspots for capacity planning, (ii) identify gaps in RF coverage, (iii) troubleshoot network anomalies, and (iv) locate users in emergency situations (E911).

Numerous techniques for localization in cellular networks exist, including methods based on time of arrival, time-difference of arrival, angle-of-arrival, cell-ID, and received signal strength fingerprinting ([5], [6], [7]). However, due to complex RF propagation phenomena (e.g., multi-path, fast fading) and the limited number of measurements provided in UMD records, model-driven techniques, such as RF triangulation, are not effective [8], [4].

On the other hand, RF fingerprinting is a *data-driven* approach to localization which makes use of the reproducibility of the received signal strength measurements. In a training phase, a fingerprint of the RF signature is empirically measured and stored for each geographical location. Then, in an online phase, the UMD channel measurements are matched to the location with the closest RF signature. However, the process of collecting training data is quite expensive and consists of two possible methods: *war-driving or crowd-sourced measurements*. War-driving involves specialized equipment placed in a vehicle and used to measure the RF signatures along roads and has 2 main challenges: (i) it requires a significant effort to collect and therefore there is poor coverage (e.g., non-road areas), and (ii) the RF measurements of the

specialized equipment are not always representative of the RF measurements for everyday users. Using crowd-sourced training data is difficult because everyday devices can only measure LTE RF signals to a few cells at a time. Thus, obtaining a detailed RF coverage map from crowd-sourced data requires a significant number of users reporting data over a long period of time.

In this work, we **leverage training data from a first-of-its-kind crowd-sourced channel measurement campaign to design the Network-Based Localization (NBL) System.** The system estimates user's locations in a 4G LTE network from their UMD records. The outline of the paper is as follows.

In Section III, we describe 4G LTE channel measurements, focusing on the parameters which will be used for localization. Notably, these include the Reference Signal Received Power (RSRP), the Received Signal Strength Indicator (RSSI), and the Propagation Distance (PD). Based on these parameters, which are reported as part of UMD to the network edge, the NBL System estimates the user's location.

In Section IV, we describe the offline training phase of the NBL System which involves building a coverage map from a crowd-sourced measurement campaign, termed GPS Tagged User Measurement Data (GUMD). The data is collected from a subset of network users, each of which has a software application installed to periodically send GPS-tagged RSRP and RSSI measurements to a centralized network server. We condense over 100TB of GUMD from over 4 million users into a coverage map, demonstrating that bivariate normal distributions can be used as a good approximation of the RF coverage in each fingerprint location.

Section V describes the online stage of the NBL System. We present an algorithm to assign weights to each location in the network based on the UMD measurements of the RSRP, RSSI, and PD. Additionally, the algorithm incorporates a pre-computed population density. Based on the assigned weighting to each grid location, the algorithm utilizes two probabilistic methods to estimate the user's location: a maximum likelihood estimator or a weighted average estimator. The algorithm is computationally quick and able to be deployed at the network edge, handling thousands of users per second.

Finally, in Section VI, we perform an extensive evaluation of the NBL System. By matching GPS records from GUMD to the UMD records collected from a national cellular operator, we obtain a data set of ≈200,000 test cases. We analyze the test cases in detail for various parameters of the system, including geographic region (rural vs. urban), coverage map resolution, and number of cell measurements in a record. Furthermore, we demonstrate the improvements in localization accuracy achieved by incorporating parameters that are not typically considered (i.e., population density estimates). We show that the NBL System can shrink the area of uncertainty around a user's location to less than 5% of the cell's coverage range. This corresponds to a median accuracy of 50m and 300m for urban and rural areas, respectively. Additionally, we describe numerous practical aspects of the system, including the deployment considerations and computation time.

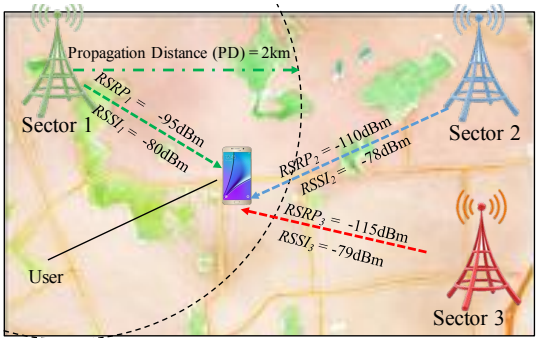


Fig. 2. Network collected UMD: The user periodically reports its measurements to the operator including RSRP, RSSI, and PD.

The main contributions of this work are:

- **The NBL System:** We design and develop the NBL System which collects and analyzes large-scale GUMD into a first-of-its-kind crowd-sourced RF coverage map. We then propose an algorithm to match UMD records composed of RSRP, RSSI, and PD measurements to coverage map locations.
- **Large-scale system evaluation, with comparable accuracy:** Using UMD from a national cellular operator, we evaluate the NBL System, showing that the localization error achieved is comparable to related works, albeit with no assumptions on user mobility. To the best of our knowledge, this is the first large-scale evaluation of a cellular localization system.
- **Practicality:** We demonstrate the practicality of the NBL System; it can be deployed at network edge locations and handle location lookups for thousands of UMD records per second. Furthermore, it can scale for national networks as it replaces extensive war-driving training measurements with crowd-sourced data.

II. Related Work

Localization in wireless networks is a well-studied field, with approaches that utilize time of arrival, time-difference of arrival, angle-of-arrival, cell-ID, and received signal strength (see [5], [6], [7] and references therein). Model driven approaches typically use geometric techniques to triangulate the user from 3 or more channel measurements to nearby access points (e.g., signal strength, angle-of-arrival). However, because we focus on *network* based localization from UMD, there is not enough information to triangulate the user [4].

Data-driven approaches to cellular localization utilize an RF fingerprint map to match a user's signal strength measurements to locations [9], [4], [5]. These techniques require training measurements, which are usually collected by placing specialized network signal measurement equipment in the back of a vehicle as it slowly traverses roads and records location and signal strength information, a process termed *war-driving*. However, war-driving does not cover all areas in the network (i.e., non-roads or major highways). Furthermore, there can be differences between the signal measured by the special equipment, and that measured by a user's smartphone. Unlike

TABLE I
COMPARISON OF THE NBL SYSTEM TO RELATED WORK IN DATA-DRIVEN LOCALIZATION FOR CELLULAR NETWORKS

Prior Work	Training Data Collection Method	Testbed Size/Location	Technology	Parameters Collected	Accuracy/Assumptions
CellSense [5]	War driving	1 rural and 2 urban testbeds in Egypt ($\approx 10\text{km}^2$)	GSM	RSSI	Median error of 27–56m, but requires >8 cell measurements.
PF^{2}S [9]	War driving	4 routes in New Jersey ($\approx 30\text{km}^2$)	3G	E_c/N_o, RSSI	Median error of 23m, but assumes a mobile user with a-priori knowledge of route and trajectory information.
Ray et. al. [4]	War driving	1 testbed in New York City ($\approx 10\text{km}^2$)	4G LTE	RSRP, RSSI	Median accuracy of ≈25m, only in urban environment for mobile users, and requires extensive per-user training for location-transition probabilities.
Network-Based Localization (NBL)	Crowd-sourced	USA-based national network ($\approx 7{,}000{,}000\text{km}^2$)	4G LTE	RSRP, RSSI, PD, Pop. Density	Reduction in localization uncertainty to 5% of the cell coverage area, corresponding to a median localization error of 50m in urban environments. Works with mobile or static users and any number of measurements.

prior work, the NBL System utilizes a first-of-its-kind channel measurement campaign based on *crowd-sourced* training data spanning the entire network of a major US network operator. We contrast the NBL System with the closest related works in Table I.

In addition, a lot of effort has recently been spent in the area of indoor wireless localization for WiFi, RFID, and Bluetooth networks [10], [11], [12]. Due to the limited covered areas, there is significantly less effort required to obtain training measurements. Additionally, these works typically assume that users provide information such as angle-of-arrival or time-difference of arrival, which are not always available in UMD.

Finally, there has been numerous works focusing on mobile trajectory tracking (e.g., [9], [4], [13], [14]). These techniques match a time-series of UMD records to a route. Thus, they only work with mobile users. The NBL System does not assume that users are moving and thus does not incorporate any trajectory tracking methods.

III. SYSTEM METHODOLOGY

In this section, we describe an overview of 4G LTE channel states as they pertain to localization [15], as well as a description the UMD collection methodology.[1] Based on the measurements included in the UMD, we then formulate the localization problem and outline the NBL System.

A. 4G LTE Background

In the 4G LTE network, basestations (also known as eNodeBs), consist of a number of cell sectors, as shown in Fig. 2. The sectors are multiplexed together spatially (using sectorized antennas) and using frequency division duplex. In this paper, we will use the terms *cell* and *sector* synonymously. Each cell establishes data sessions with users using Orthogonal Frequency Division Multiple Access (OFDMA). In OFDMA, users are allocated radio resources that span time and frequency dimensions.

For cell selection and handover, each user must estimate its channel quality to neighboring cells. To enable this, every downlink OFDMA frame contains a set of reference signals (typically 4). The reference signals are in predefined locations within the OFDMA time-frequency grid such that they capture a range of frequencies and the interference of reference signals between neighboring cells is minimized. Based on these reference signals, the following parameters are computed:

Definition 1: The average received power from the reference signals in an OFDMA frame to a neighboring cell, as computed by the user, is termed the **Reference Signal Received Power (RSRP)**.

As a measure of power, its units are in dBm and the reporting range is between -140dBm and -44dBm. Although it is not directly a measure of channel quality as it does not incorporate noise, a strong RSRP often implies the user is close to the cell center and therefore has a strong channel quality.

Definition 2: The total received power in the frequency band including power from serving and non serving cells, adjacent channel interference, and thermal noise, is computed by the user and termed the **Received Signal Strength Indicator (RSSI)**.

In practice, the RSSI measurement is used to compute the *quality* of the RSRP measurement, as compared to other interfering sources. Specifically, the Reference Signal Received Quality (RSRQ), is computed as,

$$RSRQ_i = \frac{RSRP_i}{RSSI} \cdot (\text{Number of Resource Blocks for Cell i}),$$

where i represents a cell-specific index. Figuratively, RSRQ is the fraction of received power from a neighboring cell to the total received power, akin to a signal to interference plus noise ratio. As it is a ratio of power measurements, its units are in dB.

The received signal strength measurements (RSRP, RSSI) are spatio-temporal random processes. The random variation as a function of time, even at a fixed location, is termed fast-fading and characterized by rapid fluctuations in the received signal strength (due mainly to multipath) [16]. The randomness due to changes in location, path loss, or shadowing, which occurs on the order of seconds, is termed slow-fading. In Section IV, we use training measurements to capture the slow-fading components of the signal strength measurements and average out the fast-fading components.

For uplink channel scheduling purposes, the cell must estimate the user's propagation delay, referred to as the *timing*

[1]All data collected as a part of this project has gone through internal legal and regulatory review, and is only used in anonymous and aggregate ways in accordance with our publicly available privacy policy. No personally identifiable information (PII) is included. The use of fine-grained location data is only used to determine cell coverage maps and only for network applications.

TABLE II
UMD RECORD TYPES

Record Name	Initiation Event	3GPP Standard
Initial Attach	MME receives the Attach Request from the cell/user	TS 23.40 v9.3.0, Section 5.3.2
Context Release	The cell detects user inactivity and requests the MME to remove the user's session	TS 23.401v8.6.0, Section 5.3.5
Paging (Downlink Data Notification)	MME receives notification that data is available for the user	TS 23.401v8.6.0, Section 5.3.4.3
Service Initiated	User initiates a data connection	TS 23.401v9.3.0, Section 5.3.4.1

advance. The serving cell measures the timing advance in units of symbol time (typically 32.6ns). The user preempts it's transmission by the timing advance, such that all user's transmissions arrive at the cell at the start of the slot. Based on the timing advance, the cell computes the propagation distance:

Definition 3: The distance that the strongest component of the cell reference signal travels to reach the user is termed the **Propagation Distance (PD)**.

In practice, the PD can have numerous errors and inaccuracies. For example, there is quantization error due to the discrete symbol time (each symbol time effectively translates to a distance of approximately 10m). Additional errors are introduced due to movement of the user, changes in the propagation environment (e.g., loss of line of sight), and drift in the user's oscillator clocks.

Although the above parameters can be computed by the user on the millisecond level, the measurements will only be logged by the network when triggered. As specified via standards, the common triggers are:

- **Event A1:** The RSRP to the serving cell becomes better than a threshold.
- **Event A2:** The RSRP to the serving cell becomes worse than a threshold.
- **Event A3:** The RSRP to a neighbor cell becomes better than an offset relative to that of the serving cell.
- **Event A4:** The RSRP to a neighbor cell becomes better than a threshold.

Note that the operator can configure (to some degree) the frequency of these events by adjusting the thresholds or hysteresis periods. The data collected from these events form the basis of the UMD.

B. User Measurement Data (UMD)

Operators collect and store network log information for network diagnostics, policy enforcement, and billing purposes. The information is passively collected, with records generated according to network events (e.g., session initiated, handovers, etc). The data collection is transparent to the user; the operator controls exactly which parameters it collects and the frequency of its collection. We term this data User Measurement Data (UMD).

In a typical deployment, UMD is collected at the cell and forwarded to a centralized network component (e.g., the Mobility Management Entity (MME)). Due to bandwidth and CPU constraints at the cell, not every parameter can be collected at high frequency and the collection methodology is largely vendor specific. In our configuration, the raw UMD is aggregated at each of the network edge servers. Often, these locations are termed National Technology Centers (NTCs). The vast amount of traffic and data collected requires that the UMD be handled at the network edge as it is too costly to transport to a central location. Thus, the UMD must be parsed at the network edge and a computationally fast localization method must be provided.

There are numerous implementations of UMD by various venders. In this work, we will exclusively study and utilize Per Call Measurement Data (PCMD) [17] provided by Alcatel Lucent cells and MMEs in the network of a major telecommunications network. As shown in Fig. 1, the cell forwards UMD over a control channel to the MME, which aggregates records from each of the cells in its domain and forwards the data to a cloud storage server.

A few of the common PCMD events are described in Table II. As shown in Fig. 3, an example user will have between 500-4000 PCMD records in a day, varying based mainly on user mobility and data consumption. Of particular interest to this work, each PCMD record contains information on the user's RSRP, RSSI, and PD to nearby cell sectors. In a given NTC, this results in approximately 3.5TB of uncompressed UMD collected each day. During the busiest hours of the day, data (in compressed form) is collected at over 100Mbps.

C. Problem Description

As indicated in Fig. 1, UMD records are streamed to the network edge. The NBL localization problem is to determine the location of the user at the time that each record is generated. We do not assume that a user is mobile; this is in contrast to many prior works which attempt to track the pattern of UMD to match a user to a route or trajectory (e.g., [4], [9]). Therefore, we focus our formulation on a single user at a single time instant.

Each record contains a set of RSRP, RSSI, and PD measurements, denoted as $\mathcal{R}, \mathcal{Q}, \mathcal{P}$, respectively. Each set is composed of a set of zero or more measurements to nearby cells, correspondingly denoted as R_i, Q_i, P_i where the cells are indexed from 1 to N. Thus, the sets are represented as

$$\mathcal{R} = \{R_1, \dots R_N\},\ \mathcal{Q} = \{Q_1, \dots Q_N\},\ \mathcal{P} = \{P_1, \dots P_N\}.$$

Each UMD record contains measurements for a small subset of the cells in the network (typically between 1–3), and the dimensionality of each measurement type need not be the same. For example, the PD is often only reported for the serving cell whereas RSRP and RSSI can be reported for multiple neighboring cells. To increase the dimensionality of the reported data, records generated within 2s of one another are joined together and treated as a single test case.

Based on the observed measurements, the problem is to estimate the user's location such that the distance between the user's true location and their estimated location is minimized. However, we note that localization error can be deceiving as

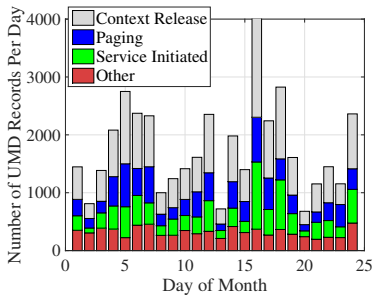


Fig. 3. Frequency of UMD records for a single user over a month.

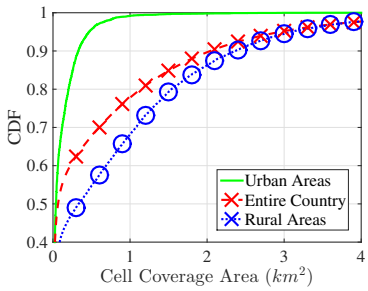


Fig. 4. Experimental distribution of cell coverage area (km^2) for 3 million cells.

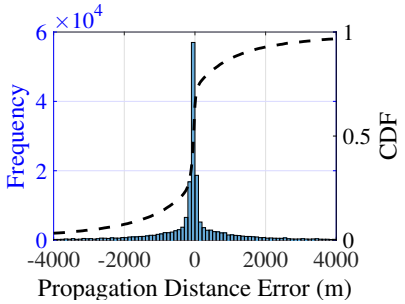


Fig. 5. Distribution of the error of 100,000 Propagation Distance (PD) measurements provided by UMD records.

TABLE III
COVERAGE MAP GRID SIZING

Zoom Level	Grid Width	Grid Area	File Size
16	411-522m	169107-274138m^2	1.0 GB
17	205-261m	42279-68534m^2	1.7 GB
18	103-131m	10570-17133m^2	2.7 GB
19	51-65m	2642-4283m^2	4.0 GB
20	26-33m	661-1071m^2	5.2 GB
21	13-16m	165-268m^2	5.7 GB

it does not account for the sizing of the cell coverage area. Hence, we also will consider a performance metric termed the *localization uncertainty*, which represents the localization error normalized by the distance covered by the cell in which the user resides. The localization uncertainty can be interpreted as the degree of uncertainty in the user's location, compared to a cell-ID based localization estimate (which corresponds to 100% uncertainty).

D. The Network-Based Localization (NBL) System

The NBL System consists of an offline phase and an online phase. In the offline phase, a coverage map is built, mapping RF measurements (e.g., RSRP, RSSI) to locations. Then, in the online phase, UMD records are streamed online to the network edge, where they are matched to coverage map locations. The matching is based on an algorithm which assigns weights to coverage map locations corresponding to their similarity (or difference) to the measurements in the UMD. We detail the offline and online phases of the system in Sections IV and V, respectively.

IV. Offline Phase: The Coverage Map

In this section, we describe the offline phase of the NBL System, consisting of collecting *crowd-sourced* training data and processing it into an *RF coverage map*.

A. GPS-Tagged UE Measurement Data (GUMD)

In this work, we generate RF coverage maps based on crowd-sourced measurement data, termed GPS-Tagged UE Measurement Data (GUMD). The data is collected from a sample of smartphone-based users on the network. Each smartphone has a proprietary application installed which periodically reports it's RF channel measurements (among other metrics) to a central server. The generated records contain data similar to the UMD records described above (e.g., including RSRP, RSSI), although a key difference is that they contain *GPS location information*.

In this work, we utilize a snapshot of all GUMD collected from an operational 4G LTE network from January 2016 through July 2016. In total, this involves over 12 TB of *compressed data* (or ≈100TB uncompressed) from over 4 million unique users. Unlike UMD, which is collected in the network by the operator, GUMD is sent over a data connection from the user to a centralized server. The data is aggregated and stored in a new file each day. The value of this data is clear; it provides location-based training information upon which we build a coverage map.

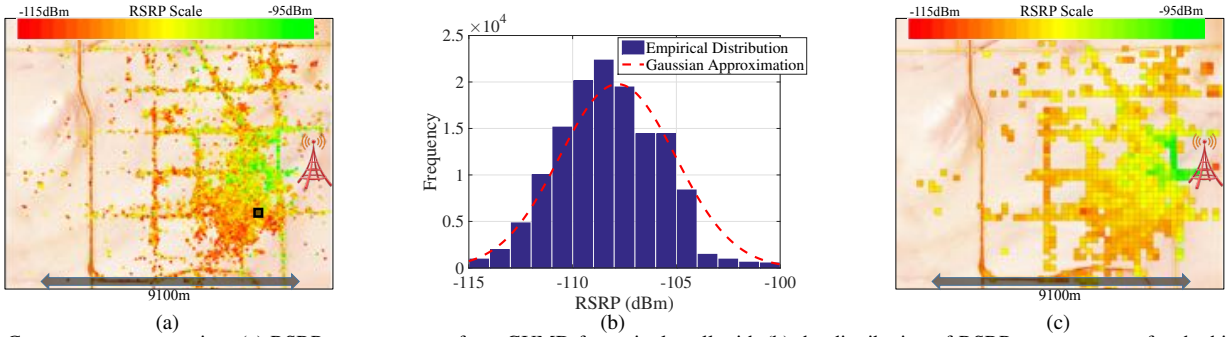
In total, we collected GUMD data from over 3 million cells. The area covered by each cell varies due to a number of factors, the primary of which is the geographical region. As shown in Fig. 4, the typical coverage area for a cell ranges up to 1 and $4\mathrm{km}^2$, for urban and rural environments, respectively.

In addition to generating coverage maps (in Section IV-B), we use the GUMD data to obtain the distribution of the error for PD measurements. That is, we correlated 100,000 PD measurements from UMD with the GPS locations from GUMD. The distribution of the error in the measurements[2] is shown in Fig. 5. The mode of the distribution is close to $0\mathrm{m}$. However, there is quite a large range in the accuracy of the PD measurements, due to multi-path propagation, processing times, clock synchronization, and numerous other affects due to vendor implementation. We store the empirical density from Fig. 5, and denote it as $f_{\mathrm{PD}}(\cdot)$.

B. Generation Methodology

An RF coverage map aggregates the raw RSRP and RSSI measurements (see Fig. 6(a)) for each grid location covered by the network. The generation of a coverage map consists of 3 stages: (i) select a representative grid of locations in the network, (ii) map training measurements to each location, and (iii) compute statistics of the RF measurements in each grid location.

[2]We compare the estimated PD from UMD to the true distance from the cell tower in GUMD.



(a) (b) (c)

Fig. 6. Coverage map generation: (a) RSRP measurements from GUMD for a single cell with (b) the distribution of RSRP measurements for the highlighted square, and (c) the mean RSRP computed in each grid location.

1) Grid: To discretize the coverage area, we overlay a grid consisting of a set of points $\mathcal{L}$. In a small covered area, it is easy to compute a set of grid locations which are uniformly spaced. However, for the large scale network evaluated in this work and due to the curvature of the earth, it is non-trivial to create a uniformly spaced grid.

To obtain a nearly-uniform set of points across the network, we utilize OpenStreetMap (OSM) tiles for our grid locations [18]. The grid locations are generated by projecting the earth into a 2-dimensional rectangle, then dividing that rectangle uniformly into a set of points. The spacing between the points is dictated by a *zoom* parameter of the OSM tiles as well as the relative location of the point to the earth's equator. For a zoom level of z, the earth is split into a grid of $2^z \times 2^z$ points. The resulting distance between grid points is presented in Table III for a few zoom levels, with the distance range presented for all points across the continental USA.[3]

Clearly, fewer grid points will result in a more condensed coverage map and thus a faster lookup. However, the grid spacing also provides the resolution for localization accuracy; that is, a sparse grid does not allow precise localization. Furthermore, with the crowd-sourced measurement campaign, there is a trade-off between grid size and *inclusion vs. dilution*. With larger grid sizes, there will be more sample data grouped together at each grid location. However, if the grid is too sparse, then each grid location will become diluted with measurements that are not representative of the RF coverage at that grid location.

2) Map: In the *map* step, each record is mapped to the nearest $l \in \mathcal{L}$. Before the mapping can occur however, we must first pre-process the records to eliminate records which have inaccurate GPS measurements. Specifically, we discard any record with a reported GPS accuracy greater than 20m. To handle the vast quantities of data, a Hadoop MapReduce job is used, taking only 1–3 hours to complete the mapping of all 100TB of GUMD.

3) Compute and Store: After each record has been assigned to a grid location, we must then simplify the information into a *coverage map*. Fig. 6(b) shows the distribution of RSRP measurements for the points falling within the grid location indicated by a black square in Fig. 6(a). As is immediately apparent, the RSRP measurements in that location appear to follow a Gaussian distribution. Thus, we approximate every grid location by storing only the first and second moments of the measurements, and modeling the distribution as Gaussian. As RSSI is closely related to RSRP, we also use this approximation to store the RSSI.

In general, this approximation closely matches the empirical measurements at each location. To quantify this, we computed the Kolmogorov–Smirnov[4] statistic for each grid location in the cell shown in Fig. 6(a), with the mean value of the RSRP in each grid shown in Fig. 6(c). Over 80% of the grid locations have Kolmogorov–Smirnov statistic that is less than 3%, implying a close match between the Gaussian approximation and the empirical distribution.

Therefore, in each grid location l, the mean value and standard deviation of the RSRP and RSSI are stored for each neighboring cell tower i, denoted as $(\mu^r_{i,l}, \sigma^r_{i,l})$ and $(\mu^q_{i,l}, \sigma^q_{i,l})$, respectively. Furthermore, from the frequency each grid location is observed in GUMD, we compute the population density at each grid location. This is computed for each location l, and stored as $\mathbb{P}(l)$. To enable fast coverage map lookups, these values are indexed by the cell tower identifier.

V. Online Phase: The Localization Algorithm

Based on the coverage map described in Section IV and the test cases described in Section III, we now present an algorithm to estimate the user's location.

A. Overall Approach

The localization algorithm is based on assigning a weight to each location $l \in \mathcal{L}$. The weight represents the degree of similarity (or difference) between the observed channel measurements, and those computed at a given grid location. We represent the weight function as $d(l, \{\mathcal{R}, \mathcal{Q}, \mathcal{P}\})$, with larger values implying a stronger match between the grid location l

[3]We note that, although RF fingerprint gridding is not a new concept [5], due to the geographical scale of this evaluation, dividing the earth into OSM tiles results in non-homogenous OSM tile sizes.

[4]The Kolmogorov-Smirnov statistic measures the difference between an empirical distribution of the RSRP and it's Gaussian approximation. The value of the statistic is the maximum value of the difference between the two CDFs at any sample value. See [19] for more details.

and the channel measurements $(\mathcal{R}, \mathcal{Q}, \mathcal{P})$. As every test case is dependent on the observed measurements, $\{\mathcal{R}, \mathcal{Q}, \mathcal{P}\}$, we simplify the notation and refer to the weight as $d(l)$. Based on the computed weights at each grid location, we present two methods to compute the user's location, denoted as l_E.

$$l_E^{\text{MLE}} = \text{argmax}_{l \in \mathcal{L}} d(l), \qquad l_E^{\text{WA}} = \frac{\sum_{l \in \mathcal{L}} d(l) \cdot l}{\sum_{l \in \mathcal{L}} d(l)}.$$

The Maximum Likelihood Estimation (MLE) of the user's location is commonly used in related works (e.g., [5]) and is denoted l_E^{MLE}. The second estimation method, l_E^{WA}, represents a Weighted Average (WA) of the user's location. In general, the MLE works well when the observed measurements closely match a single grid location. However, in cases where the observed measurements are close to a number of nearby grid locations, the WA estimator provides a tradeoff by proportionally averaging each grid location by its weight.

B. Computation of Weight Function, $d()$

Throughout this work we will compute $d()$ as the conditional probability that a user is in location l, given the observed measurements:

$$d(l) = \mathbb{P}(\mathcal{R} \cap \mathcal{Q} \cap \mathcal{P} | l) \cdot \mathbb{P}(l), \tag{1}$$

$$= \mathbb{P}(\mathcal{R}|l) \cdot \mathbb{P}(\mathcal{Q}|l) \cdot \mathbb{P}(\mathcal{P}|l) \cdot \mathbb{P}(l), \tag{2}$$

where Eqn. (2) stems from assuming independence from each of the channel measurement events. See Section III for a description of the UMD measurements. We compute the probability of each measurement observation in a given grid location l as follows:

$$\mathbb{P}(\mathcal{R}|l) = \prod_{i=1}^{N} \left(\frac{\sum_{r \in R_i} \mathbb{P}(r|l)}{|R_i|} \right), \tag{3}$$

$$\text{where } \mathbb{P}(r|l) = 2\mathbb{P}(\mathcal{N} > \left| \frac{\mu_{i,l}^r - r}{\sigma_{i,l}^r} \right|), \tag{4}$$

$$\mathbb{P}(\mathcal{Q}|l) = \prod_{i=1}^{N} \left(\frac{\sum_{q \in Q_i} \mathbb{P}(q|l)}{|R_i|} \right), \tag{5}$$

$$\text{where } \mathbb{P}(q|l) = 2\mathbb{P}(\mathcal{N} > \left| \frac{\mu_{i,l}^q - r}{\sigma_{i,l}^q} \right|), \tag{6}$$

$$\mathbb{P}(\mathcal{P}|l) = \prod_{i=1}^{N} \left(\frac{\sum_{p \in P_i} f_{\text{PD}}(\text{dist}(i,l) - p)}{|P_i|} \right). \tag{7}$$

Equations (3) and (5) compute the product of the likelihood that each RSRP and RSSI measurement, respectively, occur for all reported cells (indexed from 1 to N) at location l. Equations (4) and (6) stem from the normal approximation of the RSRP and RSSI stored in each grid location. They compute the liklihood of an observed measurement r as the probability that an instance of a normal random variable $\mathcal{N}$ with standard deviations σ is at least $|\mu - r|$ from its mean, where μ, σ represent the training value stored at grid location l. Equation (7) represents the weighting according to the propagation delay distribution from Section IV-B3.

Fig. 7. The locations of 200,000 UMD test cases, collected at one network edge site.

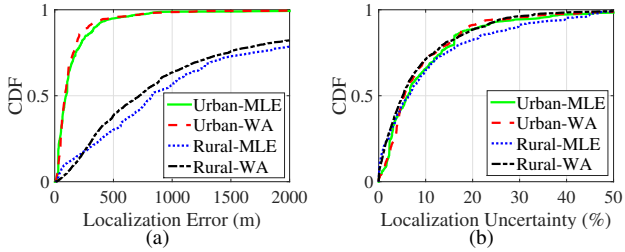


Fig. 8. NBL system experimental performance evaluation of Weighted Average (WA) vs. Maximum Likelihood Estimation (MLE) methods in rural vs. urban environments: CDF of (a) the localization error and (b) the localization uncertainty (%).

VI. System Evaluations

Using UMD test cases from a national 4G LTE network (described in Section III-C), we now evaluate the NBL System, which is composed of coverage maps (described in Section IV) and localization algorithms (described in Section V).

We study the sensitivity of the NBL System to numerous parameters including cell coverage area (e.g., rural vs. urban), coverage map resolution, and number of RF measurements. Finally, we demonstrate the improvements achieved by the NBL System from incorporating location density as well using crowd-sourced data instead of RF propagation models. Unless stated otherwise, we utilize coverage maps generated from the GUMD data with a zoom level of 19.

A. Test Case Generation

The UMD data is collected from a single network edge site and covers portions of the midwestern and northeastern USA as shown in Fig. 7. In total, we consider nearly 200,000 test cases from over 1,000 users. Each test case represents a set of UMD records generated within 2s of one another for a single user. To obtain ground truth location information, we correlate each test case with GUMD records, storing the location if a GUMD record was generated at approximately the same time (within 2s). However, the GPS locations from the GUMD may not always be accurate. For example, the UMD and GUMD systems are not perfectly synchronized, and thus aligning records by time stamps can result in the GPS locations being off by a few seconds. In addition, since the

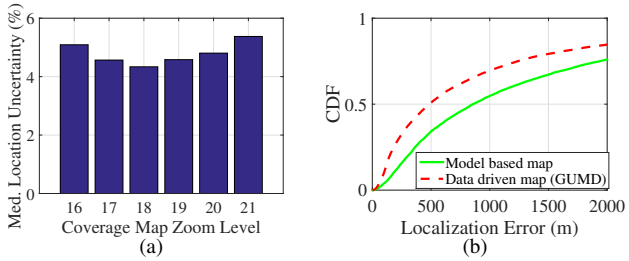


Fig. 9. NBL System experimental performance evaluation of coverage maps: (a) the median localization uncertainty for data-driven coverage maps with varying zoom levels, and (b) the distribution of localization error for a data-driven coverage map (zoom 19) compared to a model-driven coverage map.

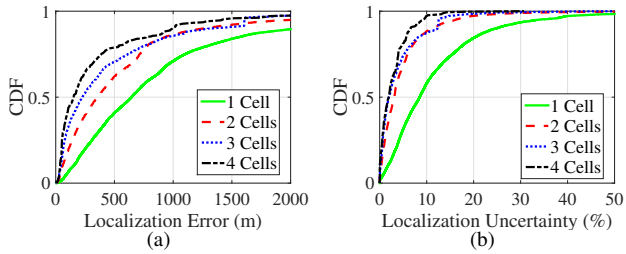


Fig. 10. NBL System experimental performance evaluation of sensitivity to number of cells reported in UMD measurements: CDF of (a) the localization error and (b) the localization uncertainty (%).

GUMD is provided through a proprietary application, we do not know the behavior of the application in scenarios with limited GPS satellite visibility. Note that related works utilize a small-scale evaluation where ground truth is collected as part of the study [9], [5], [4]. Hence, we believe that the accuracy of the NBL System may be higher than reported.

B. Localization Error vs. Cell Coverage Area

Fig. 8(a) shows the distribution of localization error achieved by the NBL System in rural vs. urban areas.[5] The localization error in urban environments (median error of ≈80m) is an order of magnitude better than in rural areas (median error of ≈750m). This follows the fact that the area covered by urban cells is much smaller than the area covered in rural cells (see Fig. 4).

Therefore, in Fig. 8(b), we present the *localization uncertainty*, which shows the localization error normalized by the cell's coverage range. With the normalized metric, the performance of the NBL System is similar in both rural *and* urban environments (each achieving a median normalized error of ≈5%). To put this into context, a typical WiFi access point with a range of 100m would have a corresponding localization error of 5m.

C. Weighted Average vs. Maximum Likelihood Estimation

Figures 8(a) and 8(b) also consider the performance of the two localization methods presented in Section V: Maximum Likelihood Estimation (MLE) and Weighted Average (WA). In practice, the WA method will have a better localization error in dense cells, as the location estimate is effectively the centroid of the most likely grid locations. However, for cells which do not have a contiguous coverage area (e.g., a body of water or uninhabited terrain in the middle of the cell's coverage), the WA method can result in poor location estimates. In the large scale evaluations, Figs. 8(a) and 8(b) show that across the entire network, the WA method has a slightly better performance. Thus, for all following experiments, we focus only on results from the WA method.

D. Coverage Map Generation

As described in Section IV, coverage maps can be generated with varying resolution. Fig. 9(a) shows the median localization uncertainty for the NBL System when using coverage maps with zoom levels between 16–21. Coverage maps which are too coarse (e.g., zoom 16 and 17) lose accuracy due to the granularity of the coverage maps. Coverage maps which are too fine (e.g., zoom 20 and 21) are limited due to lack of data collected in each grid location. Therefore, we found that a zoom level of 18–19 has the best performance when generating coverage maps from GUMD.

E. Propagation Models

As a means of comparison, Fig. 9(b) compares the performance of the NBL System using data-driven coverage maps from GUMD with a model-driven coverage map. The model-driven coverage maps are generated based on RF propagation and ray-tracing methods. For this method, we utilize a proprietary 3rd party solution developed by Forsk [20]. This method uses prior knowledge of the cell specifications, including frequency band, transmit power, transmission angle, and antenna characteristics to generate estimates of the RF coverage of each cell. The method is based on traditional ray-tracing and RF propagation models [16]. Furthermore, it takes into account known obstacles such as buildings and geographical landmarks (i.e., hills, forests, etc.). As indicated in Fig. 9(b), the data-driven approach has results in improvements in the localization error ranging from 300-1000m.

F. Sensitivity to Number of Cell Measurements

Each RF measurement in a UMD record provides added information for the NBL System to incorporate into the location estimate. Thus, as the number of measurements increases, so too does the accuracy of the NBL System. In test cases with 3 or more measurements, the NBL System achieves a median error of 50m and 300m for urban and rural areas, respectively.

Figures 10(a) and 10(b) show the performance of the NBL System for all test cases separated by the number of cell measurements in each test case. Most importantly, we note that even when the number of UMD measurements is small (less than 3), the NBL System is capable of estimating a user's location. This is in contrast to triangulation methods which require 3 or more measurements.

With limited measurements, the NBL System is able to achieve relatively-accurate localization due to the fact that the NBL System incorporates the location density and propagation distance. Figures 11(a) and 11(b) show the performance of the NBL System for test cases with only 1 or 2 cell measurements,

[5]Rural areas represent test cases generated in Pennsylvania and urban areas represent test cases generated in the New York City area.

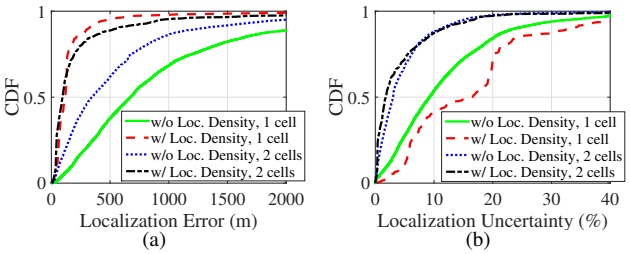


Fig. 11. NBL System experimental performance evaluation of importance of pre-computed location densities: CDF of (a) the localization error and (b) the localization uncertainty (%) for test cases with 1 or 2 cells with and without incorporating the pre-computed location densities.

with and without using the location densities. In the 1 cell case, the location densities improve the median performance by nearly 500m and 7% localization uncertainty. The benefit provided by incorporating location densities diminishes as the number of cell measurements increase. Hence, in future work, we will consider variations in the NBL System which only incorporate location density when the number of cell measurements is small.

G. Practical Considerations

In practice, the online phase of the NBL System will be implemented and deployed at a network edge location. The system must be capable of localizing a user as the data is streamed from the network. Each location lookup is independent and thus the NBL System is capable of being deployed in parallel (e.g., via Hadoop).

Each location lookup requires finding the coverage map information for every cell reported in the UMD records and assigning weights to each of the locations in the coverage area of the reported cells. In our prototype implementation written in R, it typically takes up to 13ms to estimate a users location. However, through simple optimizations such as caching coverage maps in memory via a hash map and utilizing a more performance-oriented language (i.e., C), we expect that the system could localize a user in under a millisecond. Based on users generating 1,000 UMD records per day (see Fig. 3), a 1ms estimate of computation time translates to 86,400 users handled per CPU core. Thus, a moderate deployment of 100 cores per network edge location would support nearly 90 million users. This demonstrates the scalability of the NBL System.

VII. Conclusion and Future Work

In this paper, we designed and developed the Network-Based Localization (NBL) System for a 4G LTE network. We leveraged a first-of-its-kind crowd-sourced channel measurement campaign into the creation of RF coverage maps. Then, we presented a localization algorithm to assign weights to coverage map locations based on their similarity to measurements in observed UMD records. We showed, via a large-scale system evaluation, that the NBL System achieves a median localization accuracy of 5% of the cells coverage range (corresponding to 50m and 300m in urban and rural areas, respectively). In addition, we analyzed the improvements resulting from incorporating location density and using crowd-sourced coverage maps instead of a propagation-model based map. Furthermore, we demonstrated the practicality of the system architecture, which can operate even when there is only 1 cell measurement.

Future work will focus on dynamic coverage map generation. That is, we will develop schemes to continually update the coverage map with crowd sourced data, assigning emphasis to more recent measurements for each grid location.

References

[1] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[2] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: key challenges for the radio-access network," *IEEE Vehic. Technol. Mag.*, vol. 8, no. 3, pp. 47–53, 2013.

[3] AT&T, Inc., "ECOMP (enhanced control, orchestration, management & policy) architecture white paper," http://about.att.com/content/dam/snrdocs/ecomp.pdf, Tech. Rep., 2016.

[4] A. Ray, S. Deb, and P. Monogioudis, "Localization of LTE measurement records with missing information," in *Proc. IEEE INFOCOM'16*, Apr. 2016.

[5] M. Ibrahim and M. Youssef, "CellSense: An accurate energy-efficient GSM positioning system," *IEEE Trans Vehic. Techn.*, vol. 61, no. 1, pp. 286–296, Jan. 2012.

[6] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Trans. Syst., Man, Cybern., Syst., Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, Nov 2007.

[7] S. S. Cherian and A. N. Rudrapatna, "LTE location technologies and delivery solutions," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 175–194, 2013.

[8] C. C. Cruz, J. R. Costa, and C. A. Fernandes, "Hybrid UHF/UWB antenna for passive indoor identification and localization systems," *IEEE Trans. Antennas Propag.*, vol. 61, no. 1, pp. 354–361, Jan 2013.

[9] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," *IEEE/ACM Trans. on Netw.*, vol. 24, no. 1, pp. 355–367, 2016.

[10] P. Mirowski, T. K. Ho, S. Yi, and M. MacDonald, "Signalslam: Simultaneous localization and mapping with mixed WiFi, bluetooth, LTE and magnetic signals," in *Proc. Indoor Positioning and Indoor Navigation (IPIN)*, Oct. 2013.

[11] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: unsupervised indoor localization," in *Proc. ACM MobiCom'12*, Aug. 2012.

[12] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: zero-effort crowdsourcing for indoor localization," in *Proc. ACM MobiCom'12*, Aug. 2012.

[13] S. C. Ergen, H. S. Tetikol, M. Kontik, R. Sevlian, R. Rajagopal, and P. Varaiya, "RSSI-fingerprinting-based mobile phone localization with route constraints," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 423–428, 2014.

[14] J. Paek, K.-H. Kim, J. P. Singh, and R. Govindan, "Energy-efficient positioning for smartphones using cell-ID sequence matching," in *Proc. ACM MobiSys'11*, June 2011.

[15] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution*. Wiley Online Library, 2015.

[16] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*. Prentice Hall PTR New Jersey, 1996, vol. 2.

[17] Alcatel-Lucent LTE, "Per call measurement data (PCMD) reference guide," Tech. Rep. Release LR15.1.L/WM9.0.0, Apr. 2015.

[18] Open Street Map, "Slippy map tilenames," http://wiki.openstreetmap.org/wiki/Slippy_map_tilenames.

[19] A. Stuart, M. G. Kendall *et al.*, *The advanced theory of statistics*. Charles Griffin, 1968.

[20] Forsk, "Atoll aster propagation model," http://www.forsk.com/atoll/aster-propagation-model/.