Un-Zipping Cellular Infrastructure Locations via User Geo-intent

Gyan Ranjan*, Zhi-Li Zhang*, Supranamaya Ranjan†, Ram Keralapura† and Joshua Robinson†
*University of Minnesota, Twin Cities, MN and † Narus Inc., Sunnyvale, CA.

***Abstract*—Despite the rapid growth in *cellular data* traffic, we know very little about the (*operational*) *cellular data service network* (CDSN) infrastructure. A key step in the process of developing any such understanding is to first understand the locations and distribution of the basestations in the CDSN infrastructure that serve as physical access points for end users for communicating with the underlying network. Such knowledge not only can provide critical insight into the the CDSN infrastructure, but can also guide the development of innovative (e.g. location-aware) services and applications. In this paper we propose a novel approach for mapping the CDSN basestation infrastructure via *(explicit) user geo-intent*. The intuition behind the proposed approach is to exploit specific geo-locations (i.e. geo-intent) contained in user queries to location-based services, and correlate them with basestation id's to geo-map the CDSN infrastructure. To investigate the validity of our approach, we employ data (RADIUS/RADA data sessions and application sessions) collected at the core IP network inside a CDSN. We develop heuristics for identifying user geo-intent and for geo-mapping the CDSN infrastructure — in particular, the basestations — and evaluate their efficacy using a subset of basestations with *ground-truth* GPS locations.**

I. Introduction

With wide adoption of smart phones and other mobile devices, there has been an unprecedented rise in services that rely on the locations of users (e.g. weather, maps, *locate-me*). Despite this tremendous growth, we know little, for example, about the topology and geographical distribution of the cellular network substrate (basestations); shrouded in secrecy by cellular service providers (CSP) for security reasons.

An ecosystem of services that *actively* collect information about the physical location of cellular substrate (basestations) and Wi-Fi hotspots [2], [4] already exists. Often, this is accomplished via "war-driving" (dedicated users) the entire region of interest and collecting geo-spatial data via a GPS-enabled smartphone, thereby associating the basestations with the geo-location of the "war-driving" user. Such an approach is both expensive and prone to the frequent churn in infrastructure - coverage expansion, mergers-and-acquisitions of CSPs etc. Our aim, amongst others, is to provide a cost effective (passive) alternative to such a "war-driving" approach.

In this paper we propose and explore a novel approach to *map the CDSN basestation infrastructure via (explicit) user geo-intent*. By *geo-intent*, we mean (explicit) geo-location information specified by users while submitting queries to certain services (e.g. weather or map services), in which they explicitly seek information regarding a specific location. Such geo-intent may be associated with the target of a user query, or the source (i.e. the user's own location). The basic intuition behind our approach is two-fold: i) mobile users often explicitly express their geo-intent when performing certain location-specific queries; and ii) more often than not, their geo-intent is *local*, namely related to a location in close vicinity to their current location, e.g. a nearby restaurant or the local weather. By correlating the user geo-intent thus expressed in location-specific queries with information regarding the CDSN infrastructure, e.g. the basestation a mobile device is currently associated with(such information may be obtained from mobile devices[1]), we can potentially geo-map the CDSN's basestation infrastructure.

To investigate whether — and to what extent — our proposed approach can help geo-map the CDSN infrastructure, we employ two sources of data collected at a link inside the (wired) backbone IP network of a CDSN. The first data source comprises of the RADIUS/RADA packet data sessions which contain the basestation id's (BSIDs) and *anonymized* user id's; the second data source is collections of application sessions which contain URLs extracted from HTTP headers and (anonymized) user id's. We first mine the URL datasets to extract location-specific services/apps in order to identify user queries that likely express explicit geo-intent. We find that the most prevalent type of geo-intent queries in our datasets are zip-code containing weather queries in which users seek weather information for the location specified by a zip-code. Hence, in this paper we focus our investigation on the efficacy of utilizing zip-codes in weather queries for geo-mapping the CDSN infrastructure.

Using the basestation with ground-truth GPS locations that also see zip-code queries (more than 20% of the total basestations contained in our datasets), we evaluate the efficacy of geo-mapping the CDSN infrastructure using zip-codes as user geo-intent. We find that we can, in general, geo-localize more than 50% of these basestations within $3-4$ km and more than 75% of them within $5-6$ km (alternatively, within one or a few neighboring zip-code areas). The granularity is higher for densely populated urban and metropolitan areas as shown in latter sections. Based on these observations we develop effective heuristics which exploit user geo-intent as well as user mobility for geo-mapping not only those basestations which see zip-code queries, but also those basestations which

[1]For example, some smart phone mobile operating systems, e.g. Window Mobile OS, provide certain APIs via which the BSID of the basestation a mobile device is associated with can be obtained.

do not, but instead are associated with users who issue geo-intent queries at a nearby basestation [2].

Admittedly, one limitation of our datasets is that they do not contain many mobile users with GPS-enabled devices. However, for a small number of users who do have GPS on their device, we can geo-map the associated basestations at finer granularity and better accuracy (within a few hundred meters to 1 km). We expect that with the increasing popularity of newer generations of GPS-enabled devices, our methodology will likely yield far better results than reported here.

The remainder of the paper is organized as follows: Section I-A discusses related work. Section II provides some background on the CDSN infrastructure and the datasets used. Section III presents our methodology for identifying and extracting geo-intent. Section IV describes the various geo-mapping heuristics employed, and the paper is concluded in V.

A. Related Work

Much of the existing work on localization in cellular networks has focused primarily on geo-locating mobile users or devices via signal strength based methods (e.g. triangulation) using known locations of cell towers (basestations). For a very recent study on this topic and related work, see [7] and the references therein. In contrast, we attempt to address the converse problem, namely, utilizing user geo-intent to map the CDSN infrastructure. The notion of user geo-intent has been proposed and studied recently in a different context, *web search*, with the goal to return search results that are more relevant to user queries. For instance, in [5], the authors analyze search queries from users, and classify them into explicit geo-intent and non-geo-intent queries. Our work adopts a similar notion of (explicit) geo-intent and applies it to geo-map the CDSN infrastructure.

II. Preliminaries and Datasets

A. CDSN Infrastructure

In the traditional layered network architecture terms, a typical (3G) cellular data service network (CDSN) infrastructure consists of a (layer-1/layer-2) cellular network substrate comprising of a large number of basestations and radio network controllers (RNCs) geographically dispersed across the entire coverage of a cellular service provider (CSP). Each basestation is uniquely identified by its *Basestation Identifier* (BSID), that contains three parts: the System Identifier (SID), Network Identifier (NID), and Cell Identifier (CID). The BSID namespace is hierarchical and has geo-physical significance. An SID spans a large geographical region (e.g. one or more states in the US), and is composed of multiple NIDs, each in turn representing a smaller geo-physical area. An NID, consists of many basestations, each covering a cell which is uniquely identified by a CID. Fig. 1(b) and (c) respectively illustrate the geo-physical clustering of five sample SIDs (represented by different shaded clusters), and five NIDs within a single SID. A database of SIDs, publicly available on the Internet [3], provides ownership (CSP) and geo-location details - usually the name of the city associated with the SID and the state in which it lies. Though coarse-grained, this database serves as a good cross-reference in our analysis.

[2]Determined by the time interval between accesses by common users.

B. Datasets

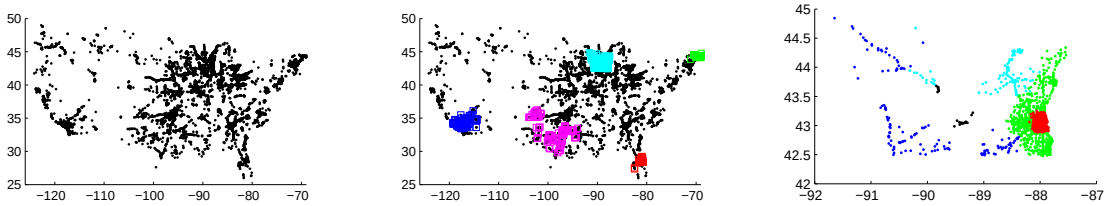
Two datasets are used in our study, which are collected at a link *inside* the core IP network of a large North American cellular 3G service provider. The first dataset (henceforth referred to as *DS-I*) was collected during a week-long period in October, 2008, representing 2 million users with 24 million packet sessions containing 110 million application sessions. The second dataset (*DS-II*) was collected over a single day in July, 2009, with 1.7 million users, 13 million packet sessions containing 147 million applications sessions. Each dataset consists of two sources of data: RADIUS/RADA *packet data sessions*, and *application sessions*. The RADIUS/RADA packet data sessions contain records of user activities such as the beginning and end times of a user's data session, the (anonymized) user id, the basestations (BSIDs) the user's mobile device is associated with during the data session etc. The application sessions records are the HTTP headers of users' Internet activities. We correlate the records from the two data-sources on the basis of the anonymized IP address in an HTTP application session, and match the HTTP timestamp such that it is between two consecutive RADA START and STOP messages, in the RADIUS/RADA packet data sessions. The URLs accessed in HTTP application sessions are extracted for identifying geo-intent queries. The BSIDs and (anonymized) user ids are extracted from the RADIUS/RADA packet data sessions. We primarily exploit the HTTP URLs, BSIDs and (anonymized) user ids, for geo-mapping the CDSN infrastructure. To verify and validate our geo-intent based mapping approach, we also utilize a collection of basestations for which we have the *ground-truth* GPS locations; this forms a representative subset of basestations spread across the US mainland (see fig. 1(a)).

III. Explicit Geo-Intent of Users

This work explores whether we can exploit "explicit geo-intent" of mobile users to learn the CDSN infrastructure, i.e. the physical locations of basestations. We define *explicit geo-intent* as location information contained in queries submitted by users to certain services (e.g. weather or map services) in which they seek information regarding a specific location. We now discuss various aspects of the explicit geo-intent in our dataset DS-I (owing to its week-long span) in the following subsections.

A. Extracting Explicit Geo-intent

We employ a set of heuristics to identify and extract geo-intent from the HTTP URLs in our datasets. Our objective is to find a set of services seen in our URL trace with a geo-intent format that can be automatically extracted, giving us a mapping between URL and the geo-intent expressed in that URL.



(a) US mainland (b) Five sample SIDs (c) Five NIDs within an SID

Fig. 1. Illustration of geo-physical clustering of BSID's at SID/SID-NID level (Ground-truth set).

Through a manual process of identifying a set of location-specific keywords, such as street or state names, zip-codes, and "GPS-like" latitude-longitude co-ordinates, that may or may not be directly provided by the (satellite) global position system (GPS) service (see §III-C for a details), we create a set of rules to perform such extraction. The output of this step are rules for extracting the embedded geo-physical identifiers in the URL string for each hostname (e.g. `www.weather.com` or `www.mapquest.com`). Through such rules (heuristics), we identify over a million URLs with geo-identifiers from DS-I and half a million from DS-II.

On closer scrutiny of the geo-identifier information contained in the extracted URLs, we find that zip-codes and "GPS-like" co-ordinates dominate the URL geo-identifier set accounting for $\approx 99\%$ of the URLs we are able to parse. Moreover, such URLs primarily belong to weather information internet hosts. In the following subsections, therefore, we focus on such URLs from the weather services to determine whether or not these URLs indeed reveal users' geo-intent.

B. Zip-codes in Weather Queries

Weather queries are obvious candidates for finding zip-code information due to the nature of online weather services. Most phones feature a weather application allowing users to enter/store the zip-codes for one or more locations of interest. Often, these queried locations represent the user's home or place of work. Therefore, the zip-codes in weather queries provide a good, though not always precise, indication of the querying user's location. We later evaluate the usefulness and accuracy of such zip-code information in our datasets for the purposes of geo-mapping the CDSN's basestation infrastructure.

Throughout this work, we convert the zip-codes in geo-intent queries to GPS-like co-ordinates as follows. The US census bureau [1] provides approximate boundaries of the zip-code tabulation areas (ZCTA) [3] encompassing each zip-code in the US. From the boundary co-ordinates for a given zip-code, we compute the *centroid* (a pair of latitude-longitude co-ordinates). In what follows, the term zip-code will invariably mean the corresponding centroid location calculated as described here.

[3]Some ZCTAs may span several zip-codes in less populous regions. As our results later show, for our purpose the ZCTAs provide sufficient accuracy.

C. GPS-like co-ordinates in Weather Queries

Next, we investigate the URLs (weather and other queries) containing GPS-like (latitude-longitude) co-ordinates. A majority of these GPS-like co-ordinates appear in the HTTP responses and not the HTTP requests. Moreover, the co-ordinates in HTTP responses show significant variance. For example, an HTTP request for weather information for zip-code = 53108, received two HTTP responses with co-ordinates $L_1 = (lat = 42.82, long = -87.93)$ and $L_2 = (lat = 42.82, long = -99.76)$. Such co-ordinates, sometimes used for displaying maps on the user device, therefore, need not reflect user geo-intent in practice and thus we do away with them (see the technical report [6] for details).

However, for a small number of URLs related to a few services, e.g. $GPSToday$ hosted by `www.geoterrestrial.com` [4], we do see GPS-co-ordinates in HTTP requests. Hereafter, we refer to this small set of GPS co-ordinates as the *GPS geo-intent* dataset and return to it briefly in §IV-A.

For the remainder of the paper, we focus on zip-code information, except where noted otherwise. We remark that our geo-mapping methodology, presented later, can incorporate GPS co-ordinates with the potential to provide greater precision as more GPS-enabled devices and services are deployed.

D. Spread of Geo-intent in the Basestation Infrastructure

To associate the geo-intent expressed in users' queries with the basestation infrastructure of the CDSN, we first need to identify and extract relevant basestation information (BSID associated with a user at the time of query). Henceforth, we say that a basestation B, *sees* a zip-code Z if at least one user queries for weather information (or any information in general) for zip-code Z while communicating with basestation B. Although the number of users expressing their explicit geo-intent is a small fraction of the overall user-base (less than 2%), the number of basestations that see at least one zip-code query is significantly large ($\approx 23\%$); and cover a representative fraction of the infrastructure (SID-NID pairs/ SIDs). Therefore, explicit geo-intent is pervasive not only in terms of geographic coverage but also in the CDSN infrastructure.

[4]GPSToday is a service that provides topographical (e.g. altitude) and weather related information at the user' current location. Hence the GPS co-ordinates contained in user queries to this service reflect *explicit user geo-intent* at the source (user) location.

Clearly, if the geo-intent of users indeed captures their geo-location, it can possibly help geo-map a significant fraction of the basestation infrastructure across wide geographies.

Moreover, we observe that the 50^{th} and 75^{th} percentiles for the ratio of unique zip-codes to the number of geo-intent queries seen at a basestation are 0.2 and 0.4 respectively; a relatively small number. This observation has important implications in the process of geo-mapping of the basestation infrastructure, as will be explored in the next subsection.

E. From Geo-intent to Geo-location

With about 23% of the basestations in our dataset seeing zip-code containing weather queries, can we use the explicit geo-intent information contained therein to geo-localize the basestations in question? Among the basestations with ground-truth GPS locations, we find that roughly 20% (a representative sample) also see zip-code queries; moreover, they span 105 SID-NID pairs across 81 SIDs. In the following we will refer to the set of such basestations ($\approx 2,500$ in all), with both the ground-truth GPS locations and associated zip-code queries, as the *ground-truth-location-&-zip-code* BSID dataset.

We now examine the relationship between the locations of the basestations and users' geo-intent (the zip-codes associated with the basestations). Given a basestation B with *known* GPS location denoted by $L_B = (lat_B, long_B)$, let $Z_B = \{Z_1, Z_2, ..., Z_k\}$ be the set of zip-codes seen at B. Let $C_{Z_i} = (lat_i, long_i)$ be the co-ordinates for the centroid associated with Z_i. We denote the distance (in km) between the basestation B and the zip-code Z_i by $\delta_i^B = dist(L_B, C_{Z_i})$ [5]. In particular, we define $\delta_{min}^B = min_{1 \le i \le k} \delta_i^B$ and $\delta_{max}^B = max_{1 \le i \le k} \delta_i^B$. Further, for basestations that are associated with multiple zip-codes ($k \ge 3$), we also compute the distance between each basestation and the most frequently queried zip-code/s associated with it (δ_*^B).

For the basestations in the *ground-truth-location-&-zip-code* set, the distance between basestation B and the closest queried zip-code (δ_{min}^B) is within $1 - 1.5$ km range for 25%, $3 - 3.5$ km for 50% and within $5 - 6$ km for 75% of basestations. We observe similar distributions for δ_*^B. In contrast, the corresponding figures for δ_{max}^B are: within 7 km for 25%, 12.5 km for 50% and 20 km for 75% of basestations. In short, while the distance between the true location of a basestation and the farthest queried zip-code seen by it can be quite hight (10's km), the closest and most frequently queried zip-codes (often same for most basestations) are often within reasonable proximity (5-6 km).

In order to incorporate varying sizes of zip-code areas (zip-code sizes vary between urban and rural areas), we use the centroid of each zip-code and perform a *Voronoi partition* of the entire US mainland [6]. Given a basestation B with known GPS location $L_B = (lat_B, long_B)$, let $\widehat{Z}_B$ be the *home* zip-code of B and h_i^B the hop-count distance between $\widehat{Z}_B$ and $Z_i \in Z_B$ in the Delaunay graph (dual of the Voronoi partition). We observe that for 90% of the basestations, $h_i^B \le 3$ for over 75% of Z_i's seen at them. We, therefore, conclude that for a large majority of basestations, a significant percentage of zip-codes queried are in and around the geo-physical neighborhood of their home zip-codes.

[5]Haversine distance with earth's mean radius = 6,371 km.

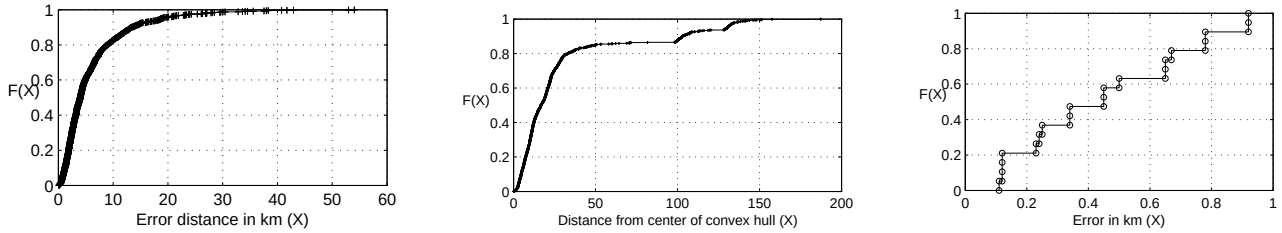
[6]Instead of partitioning the US mainland in terms of the ZCTA boundaries using data from [1], we use the Voronoi partitions for ease of analysis and computation.

IV. Geo-mapping the Basestation Infrastructure

Based on the analysis and observations made in the previous section, we now present some heuristics to geo-map the basestation infrastructure. For evaluating the accuracy of each suggested heuristic, we make use of dataset DS-II in this section.

Direct Geo-mapping via Geo-Intent. For those basestations which see at least one zip-code containing weather query, we directly geo-localize them using explicit geo-intent by means of the following heuristics. Given a basestation B, let $Z_B = \{Z_1, Z_2, ..., Z_k\}$ be the set of valid zip-codes queried by its users $U_B = \{U_1, U_2, ..., U_l\}$. The first heuristic, the *Majority Voting* (MV) scheme, selects the most probable location (or locations) from all possible zip-code locations (Z_i)'s as follows: Each user $U_i \in U_B$ has one simple vote. Recall that a given user U_i may query the same zip-code Z_j multiple times. In order to prevent such frequent voters from skewing the vote count, we permit a user to vote only once. Also, a given user U_i may possibly query multiple zip-codes from the set Z_B. In such cases, we split the simple vote of U_i proportionally among all the zip-codes s/he queries. For example, if U_i queries zip-code Z_1 thrice and Z_2 twice, Z_1 receives 0.6 vote and Z_2 receives 0.4 vote from U_i. The winner of the election, i.e. the zip-code receiving most votes, is chosen as the most probable geo-location for B. When there are multiple winners (ties), all of them are chosen as equally probable locations. Fig. 2(a) shows the error incurred (compared to the ground-truth) in geo-mapping approximately 3,500 basestations using the MV-scheme. We observe that for 75% basestations thus geo-mapped, the error is within $5-7$ km range, quite similar to what we observed for DS-I in §III-E.

Indirect Geo-mapping based on User Mobility. The direct geo-mapping via geo-intent helps geo-localize around 20% of the basestations in our datasets. To map other basestations, those not mapped during direct geo-mapping due to lack of geo-intent queries, we exploit user movement. To do so, we introduce the *basestation-user-mobility graph*, $\mathcal{G}_M$, where the vertices are the basestations (BSIDs) and an edge $e = (B_i, B_j)$ is introduced between two vertices B_i and B_j if at least one user accesses both of them (regardless of order) within a short interval of time ΔT (say 5 minutes). Given $\mathcal{G}_M$ thus defined, let $\mathcal{B}_{mapped}$ denote the set of basestations geo-located via the direct geo-mapping heuristic described above, and $\mathcal{B}_{unmapped}$ be the set of *unmapped* basestations. For each basestation $B \in \mathcal{B}_{unmapped}$, if it is connected to some basestation $\overline{B} \in \mathcal{B}_{mapped}$ via some paths, we define $h(B, \overline{B})$ as the shortest path distance (hop-count) from B to $\overline{B}$. Then, let $h(B, \mathcal{B}_{mapped}) = \min_{\overline{B} \in \mathcal{B}_{mapped}} h(B, \overline{B})$.



(a) Dist. Err. (MV-scheme) (b) Dist. Err. Indirect (c) Dist. Err. Geo-Terr. GPS set

Fig. 2. Error in geo-mapping compared to the ground-truth set.

Note that $h(B, \mathcal{B}_{mapped}) = \infty$ if B is not connected to any $\overline{B} \in \mathcal{B}_{mapped}$.

In our datasets, we have about 22% basestations in $\mathcal{B}_{unmapped}$ that are connected to at least one basestation in $\mathcal{B}_{mapped}$ at a 1-hop distance (i.e. $h(B, \mathcal{B}_{mapped}) = 1$). Hence, we geo-localize them first by exploiting their connectivities to the basestations in $\mathcal{B}_{mapped}$. The challenge here is to map the connectivity in $\mathcal{G}_M$ to geo-locations or zip-code neighborhoods in the Delaunay graph of US zip-codes. For any $B \in \mathcal{B}_{unmapped}$ such that $h(B, \mathcal{B}_{mapped}) \leq d$ such that it is connected to at least s basestations in $\mathcal{B}_{mapped}$ that are at most d hops away from B, we geo-localize B by constructing a connected zip-code neighborhood in the Delaunay graph of zip-codes. Let $N_d(B)$ be the set of home zip-codes of the (directly mapped) basestations $\overline{B}$'s in $\mathcal{B}_{mapped}$ that are at most d-hops away from B (note that $|N_d(B)| \geq s$). Using the centroids of $\widehat{Z}_{\overline{B}}$'s, we construct a convex hull H_B, covering all $\widehat{Z}_{\overline{B}}$'s, as the most probable geo-location for B.

Fig. 2(b) shows the error incurred (compared to the ground-truth) in geo-mapping approximately 4,000 basestations using the user-mobility graph. Here we observe relatively higher error rates mostly in rural areas where the geo-physical expanse of zip-codes is much larger and therefore an edge in the Delaunay graph between adjacent zip-codes may cover tens of kilometers. Despite this, the mapping accuracy is much higher (to within a city) than the county level information available in the SID database [3].

A. Geo-mapping using GPS Geo-intent

Lastly, we use the small *GPS geo-intent* dataset discussed in section §III-C to illustrate the efficacy of our approach when GPS-based geo-intent (in particular, when the GPS co-ordinates are associated with the source (user) locations of the geo-intent. We extend the direct geo-mapping heuristics from §IV to the case of GPS co-ordinates, and apply tessellation and density estimation to geo-localize basestations by computing a (small) neighborhood area (rectangular cell) as their most probable locations. Due to the space limitation, the details are omitted. Fig. 2(c) shows the mean distance (error) between the ground-truth and the inferred (centroid) locations of the two dozen basestations in the small GPS geo-intent dataset (and for which we have the ground-truth locations). We see that the overall accuracy is within 0.5 - 1 km. Hence we believe that with the increasing popularity of newer generations of GPS-enabled smart phones and location-aware services, geo-mapping based on user geo-intent will yield more accurate results than what can be obtained using zip-codes alone.

V. Conclusion and Future Work

In this paper we put forth a novel approach for mapping the CDSN basestation infrastructure via *(explicit) user geo-intent*, which circumvents the handicaps plaguing conventional approaches (e.g. war-driving). We developed heuristics for identifying user geo-intent to geo-map the basestations and evaluated their efficacy using a subset of basestations with known *ground-truth* GPS locations. Using zip-codes contained in user weather queries, we demonstrated that a large portion of basestations can be geo-mapped within a $3.5-6.0$ km range in general, within $1.5-2$ km range in densely populated urban areas and often within 1 km in large metro-areas.

Given the exponential growth in cellular data traffic, we believe that mapping the CDSN infrastructure is a critical step in understanding how to best expand and evolve the CDSN infrastructure to better meet growing user demands, and to guide the development and deployment of innovative location-aware services and applications that cater to mobile users and devices. Our study is only an initial step along this direction, and much additional research is still sorely needed.

VI. Acknowledgment

We are thankful to Ionut Trestian, Aleksandar Kuzmanovic and Antonio Nucci for participating in early discussions on this work. The first two authors of this paper were supported in part by the NSF grants CNS-0905037, CNS-1017647 and CNS-1017092 and the DTRA grant HDTRA1-09-1-0050.

References

[1] *http://www.census.gov*.
[2] *http://www.navizon.com*.
[3] *http://www.roamingzone.com*.
[4] *http://www.skyhook.com*.
[5] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, *Analysis of geographic queries in a search engine log*, Proc. of LocWeb (2008).
[6] G. Ranjan, Z. L. Zhang, S. Ranjan, R. Keralapura, and J. Robinson, *Mapping cellular data service network infrastructure via geo-intent inference*, Tech. Report, http://www-users.cs.umn.edu/∼granjan (2010).
[7] H. Zang, F. Baccelli, and J. C. Bolot, *Bayesian inference for localization in cellular networks*, Proc. of IEEE INFOCOM 2010, March 2010.