Mapping Cellular Data Service Network Infrastructure via Geo-intent Inference

Gyan Ranjan*, Ram Keralapura†, Supranamaya Ranjan†, Joshua Robinson† and Zhi-Li Zhang*
*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. For example, how are the core IP network elements distributed over the cellular network substrate such as basestations? Such knowledge and understanding not only can provide critical insight into the evolution of the CDSN infrastructure, but can also guide the development of innovative (e.g. location-aware) services and applications. In this paper we propose and explore a novel approach for mapping the CDSN 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 and gateway IP addresses 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 IP NAS gateways — and evaluate their efficacy using a subset of basestations with *ground-truth* GPS locations. Our findings not only shed useful light on the CDSN infrastructure, but also have implications in the design of effective location-aware services and applications.

1. INTRODUCTION

With wide adoption of smart phones and other mobile devices, *cellular data* traffic has grown tremendously in the past few years. This growth will be further precipitated by the increasing popularity of newer generations of smart-phones and mobile devices such as iPhones, Android phones and iPads. As in the case of wireline services, cellular data traffic will likely surpass the voice traffic in the not-so-distant future. Despite this tremendous growth, there have been relatively few studies on the (*operational*) *cellular data service network* (CDSN) infrastructure. Apart from various articles, papers and documentations on the architectural design and component engineering (e.g. 3G network standards), we know little, for example, about the topology and geographical distribution of IP network elements over the cellular network substrate such as basestations. The challenges in conducting *measurement-based* mapping of operational cellular data service networks (CDSNs) can be attributed, in part, to the fact that these networks are generally "closed", unlike most of the traditional Internet infrastructure. In other words, active probing (e.g. traceroute) from outside typically elicits no response from the internal network elements of a CDSN. Conducting active probing from mobile devices in general does not help much either, as IP addresses assigned to the end users' devices are often private IP addresses [3]: the only IP addresses visible to the outside world are the IP addresses of exit routers of the CDSNs. With the rapid growth in cellular data traffic, gaining a better understanding of the CDSN infrastructure — especially the geo-spatial relationships of the IP network overlaid on top of the cellular (basestation) network substrate — is imperative. Such understanding can not only provide insights into the evolution and expansion of existing (and future) CDSNs, but also help guide the development and deployment of innovative location-aware services and applications that cater to mobile users (see more discussion on this in a latter part of this section).

In this paper we propose and explore a novel approach to *map the CDSN 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) their explicit geo-intent is often *local*, namely related to a location in close vicinity of their current location, e.g. a restaurant nearby or the local weather. Such queries will occur more frequently as

more users adopt GPS-enabled smart phone and utilize location-based services or apps on their mobile devices. By correlating the user geo-intent expressed in location-specific queries with information regarding the CDSN infrastructure, e.g. the basestation a mobile device is currently associated with or the (first-hop) IP gateway address (such information may be obtained from mobile devices[1]), we can geo-map the CDSN 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. Two datasets (containing data from both sources), collected roughly ten months apart, are used for our study. For a subset of BSIDs, we also have the *ground-truth* GPS locations. We start with a list of geographical identifiers (e.g. zip-code, street, city, state, GPS coordinates etc.) and 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. In other words, zip-codes in these queries represent explicit user geo-intent (target locations of interest). Hence, in this paper we focus our investigation on the efficacy of utilizing zip-codes in weather queries for geo-mapping the CDSN infrastructure.

While zip-code containing weather queries represent a small percentage of all URLs, the collection of basestations seeing such zip-code queries (i.e. around which such queries originate) constitute more than 20% of the total basestations contained in our datasets. Using the basestation with ground-truth GPS locations which also see zip-code queries, we evaluate the efficacy of geo-mapping the CDSN infrastructure using zip-codes as user geo-intent[2]. We find that we can geo-localize more than 50% of these basestations within 3.9 km and more than 75% of them within 6.1 km (alternatively, within one or a few neighboring zip-code areas). The accuracy and coverage vary with the population density: within large metropolitan areas in the US, we can improve the accuracy to within $1.5-2$ km for most basestations due to smaller zip-code areas and more users; while for rural areas (sparsely populated) or highway corridors (high user mobility), the accuracy is poorer due to large zip-code areas and/or noisy, non-local zip-codes queried for by users. 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 do not see any zip-code queries but instead are associated with users who issue geo-intent queries at a neighboring basestation within a short period of time. Finally, we extract the IP addresses (of mobile devices, NAS gateways, IP home-agents etc.) in the datasets and study the geo-spatial distribution of IP network elements within the CDSN infrastructure, using the inferred (and ground-truth) locations of the basestations. The results and their implications are discussed below.

Our findings not only confirm the validity of our proposed approach for geo-mapping the CDSN infrastructure using user geo-intent, but also shed light on some interesting aspects of the CDSN infrastructure, especially the geo-spatial relation between IP network elements (NAS gateways and IP home-agents) and the basestation substrate. In contrast to the number of basestations (more than 80,000 in our datasets), the number of NAS gateways and home-agents is far smaller (less than 100). Further, each NAS gateway (or IP home-agent) covers a large geographical region (typically spanning multiple states in US) and a large number of basestations (spanning multiple SIDs). Multiple gateways or home-agents may also cover similar regions, likely for load-balancing and reliability. Our observations have important implications. For example, with the rapid increase in cellular data traffic (for instance, compared with the first dataset, we see a multi-fold increase in diurnal traffic volume in the second dataset collected ten months later), IP networks within the CDSN infrastructure are likely to — and need to — undergo drastic expansion and evolution to meet the growing user demand. Our results also provide a plausible explanation why attempting to geo-localize mobile users/devices based on externally visible IP addresses (of exit routers) does not yield reliable results (cf. [3]), and may be futile apart from perhaps geo-localizing within a large geographical region. In a similar vein, deploying "location-aware" content distribution services outside CDSNs, which attempt to reduce latency by distributing content locally, may not be very effective at present. To better cater to mobile users in fine-grained geography (e.g. within a state or metropolitan area), such services may have to be deployed inside CDSNs.

[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, the gateway IP address as well as the IP address assigned to the mobile device can be obtained.

[2]Thus the accuracy and granularity of geo-mapping are roughly at the level of zip-codes which is a considerable improvement over, for instance, some publicly available information about basestations, e.g. the SID database [2] which provides geo-location of base-stations at the granularity of state(s) in the U.S.

Last but not the least, we remark that one limitation of our datasets is that it does not contain many mobile users with GPS-enabled devices; thus we identify only a small set of geo-intent queries (for a couple of location-based services) which contain GPS coordinates coming from GPS-enabled mobile devices. Using this small set of GPS coordinate geo-intent data, we show that we can geo-map the associated basestations with far finer granularity and better accuracy (within a few hundred meters to 1 km). With the increasing popularity of newer generations of GPS-enabled smart phones and mobile devices, we expect our methodology to yield better results than those reported here.

The remainder of the paper is organized as follows. In §1.1 we will briefly discuss the related work. In §2 we provide some background on the CDSN infrastructure and describe the datasets. In §3 we present our methodology for identifying and extracting geo-intent, and in §4 we investigate the efficacy of geo-localizing basestations using zip-code geo-intent. We develop several geo-mapping heuristics in §5. In §6 we study the geo-spatial distributions of the IP network elements within the CDSN infrastructure, and the paper is concluded in §7.

1.1 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 [9] and the references therein. In contrast, we attempt to address the problem the other way around, 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 [4], the authors analyze search queries from users, and classify them into explicit geo-intent and non-geo-intent queries.

In [8], the authors go one step further to extract (implicit) geographical information that can plausibly identify users' locations. Our work adopts a similar notion of (explicit) geo-intent and applies it to geo-map the CDSN infrastructure.

Trestian et al [7] correlate user-location (at the granularity of basestations) and application interests over time. In their analysis of user mobility patterns, they find that many users tend to move around one or a few location "hot-spots" (e.g. residence, office, or a coffee-shop). This finding indicates that a majority of users' geo-intent is likely local to their locations. The study in [3] cited earlier collects the IP addresses assigned to mobile devices as well as the IP addresses (likely those of exit routers or NAS gateways) which appear as source IP addresses in the queries sent to a web server under the authors' control, and uses them to locate mobile devices by geo-localizing the IP addresses.; only to find that the geo-mapping results using these IP addresses are very coarse-grained and often unreliable. Our study shows that the core IP network is "sparsely" distributed over the dense and geographically dispersed (see fig. 1(a)) cellular network substrate, thus providing a plausible explanation for their findings.

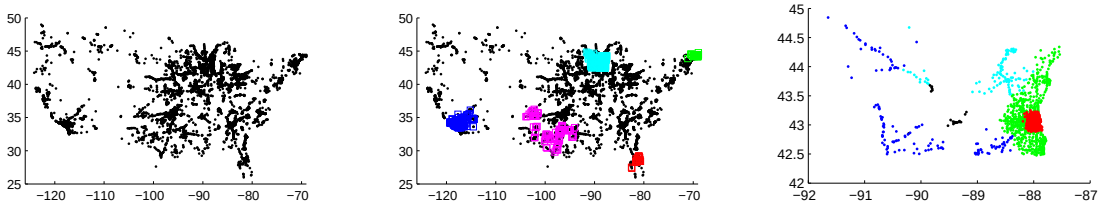
2. PRELIMINARIES AND DATASETS

2.1 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 and an IP data core network overlaid on top. The cellular network substrate comprises 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), which 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 representing a smaller geo-physical area. An NID, in turn, 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.

SIDs are allocated to CSPs by the International Forum on ANSI-41 Standards Technology (IFAST) based on territories. A database for SIDs, publicly available on the Internet [2], provides ownership and geo-location (coarse-grained) details. A typical record in this database has five attributes: a decimal value representing the SID, the city associated with the SID (usually the name of the most populous city), the state in which the city lies, name of the CSP to whom the SID has been allocated, and the operational frequency band. Though coarse-grained, the database serves as a good cross-reference in our analysis.

The IP network of a CDSN typically consists of IP gateways (usually referred to as *network access servers* or NAS gateways) through which data from/to mobile devices enters/leaves the IP network, (IP) home agents (for user registration and mobile IP routing), and other standard network elements such as routers, DHCP servers, DNS servers, and so forth. The IP



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

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

Table 1: User and traffic volume statistics.

Dataset (time)	Users	Duration	#Pkt.&App. Sess.
I (Oct 2008)	2 M	7 days	24 M & 110 M
II (Jul 2009)	1.7 M	1 day	13 M & 147 M

network also includes a number of RADIUS/RADA [3] servers for authenticating users, and for logging user data access activities for billing and accounting purposes.

2.2 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 *Dataset I*) was collected during a week-long period in October, 2008, and the second dataset (*Dataset II*) was collected over a single day in July, 2009. Table 1 summarizes overall statistics regarding Datasets I and II. Both datasets are *anonymized* packet traces. 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. Recall from fig. 1(a), the basestations in our ground-truth set are widely distributed across the US mainland and provide an extensive and representative set for verifying and validating the results obtained in our study.

[3]RADIUS stands for the Remote Authentication Dial In User Service protocol [6, 5], and RADA stands for the Radius Authenticated Device Access protocol. Both are used to provide centralized *Authentication, Authorization, and Accounting* (AAA) management.

3. 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 and the IP data network elements. 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. Such geo-intent in user queries may either be associated with the current (source) location of a user (e.g. *locate-me* type of features) or her target location of user (e.g. weather lookups for a region of interest).

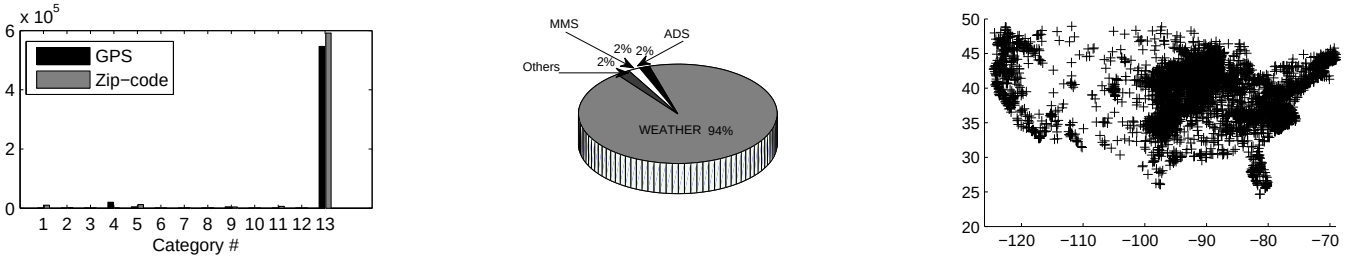
One of the greatest challenges faced in this approach is that the geo-intent expressed in a user's query is encoded in a format meaningful for specific services and therefore varies from one service to the other. To address this issue, careful service-specific analysis is required to extract relevant explicit geo-intent from user queries. In §3.1, we describe our heuristics for doing this. Next, in §3.2, we focus on the most dominant type, namely *zip-codes contained in weather-related queries*, which are primarily associated with the target locations of users' geo-intent. Lastly, in §3.3, we discuss the relevance of GPS-like information observed in our datasets and identify the cases when it is relevant and useful.

3.1 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. Through a manual process of identifying a set of location-specific keywords,

Table 2: Web services and sample URLs with geo-physical identifiers in Dataset I.

Hostname	Geo-physical identifiers in URL	# of URLs
pv3.wirelessaccuweather.com	zip=54940&city=Fremont&state=Wi&country_code=US	510,170
mapserver.weather.com	lat=43.45&long=–88.63	273,061
maps.google.com	q=starbucks&near=Oconomowoc	9,631
addshuffle.com	zip=53946&cntry=US	6,434
geo.yahoo.com	lat=42.97&lon=–88.09	3,519



(a) URL count per category (b) Combined count (%) per category (c) Zip-code geo-intent (weather)

Figure 2: Dominance and geo-physical expanse of weather (category 13) related queries in Dataset I.

such as street or state names, zip-codes, and "GPS-like" coordinates [4], 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`). Table 2 shows some examples of services and the associated geo-identifier formats. Through such rules (heuristics), we identify over a million URLs with geo-identifiers from Dataset I and half a million from Dataset II.

We further analyze the geo-identifier information contained in the extracted URLs to understand the variety of geo-identifiers they contain. We find that zip-codes and "GPS" coordinates dominate the set of URLs with geo-identifiers (in $\approx 99\%$ of the URLs we are able to parse). Henceforth, we focus only on such URLs. Furthermore, to understand better the services associated with such URLs (with zip-code and GPS-like geo-identifiers), we classify each service (based on hostname) into 13 different categories: *1-Ads, 2-books, 3-dating, 4-maps, 5-MMS, 6-music, 7-news, 8-photo, 9-search, 10-toolbars, 11-trading, 12-video and 13-weather*. Fig. 2(a) shows the number of URLs in each application category separately for two types of geo-identifiers: zip-codes and GPS-like coordinates. We see that weather services constitute the most dominant category accounting for about 94% URLs with either a zip-code or a pair of GPS-like coordinates (see Fig. 2(b)).

In the following subsections we analyze the geo-identifiers contained in the URLs in weather category to determine whether or not these URLs indeed reveal the user geo-intent.

3.2 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 the zip-codes for one or more locations of interest. Quite 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 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 infrastructure.

In this work, we convert the zip-codes contained in geo-intent queries in terms of a GPS-like coordinate as follows. The US census bureau [1] provides GPS-like coordinates which delineate the approximate boundaries of the zip-code tabulation area (ZCTA)[5] encompassing all the zip-codes in the US. Using such boundary coordinates for a given zip-code, we compute the *centroid* (a pair of GPS-like coordinates). In the remainder of this paper the term zip-code will be used exclusively to mean the corresponding centroid location calculated as described here. Fig. 2(c) shows the geographical distribution of the zip-codes (centroids) contained in all the weather queries in our datasets. We see that the set of zip-codes in the explicit geo-intent of users pervades nearly all parts of the US mainland.

[4]In this paper we refer to any pair of latitude-longitude coordinates as "GPS" coordinates, although in fact many of these may not be directly provided by the (satellite) global position system (GPS) service. See §3.3 for a detailed discussion.

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

3.3 GPS-like Coordinates in Weather and Other Queries vs. True Geo-Intent

Next, we investigate the URLs containing GPS-like (latitude-longitude) coordinates. As shown in Fig. 2, the weather category also contains an (almost) equal number of URLs with GPS-like, latitude-longitude coordinates. A majority of these GPS-like coordinates appear in the HTTP responses and not the HTTP requests. Further inspection reveals that these coordinates do not directly reflect the geo-intent of users, and show significant variance (see table 3). However we do observe a few services, e.g. *GPSToday* hosted by `www.geoterrestrial.com`, where the GPS coordinates contained in user queries submitted to these services do reflect *true* geo-intent[6]. Unfortunately, it represents a very small fraction of queries in our datasets. Hereafter we refer to this small set of GPS coordinates as the *GPS geo-intent* dataset.

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 is also able to incorporate GPS coordinates and has the potential to provided greater precision as more devices and services, which use the capabilities of GPS-enabled smart-phones, are deployed.

4. FROM USER GEO-INTENT TO GEO-LOCATIONS IN THE CDSN

In this section we correlate the zip-codes extracted from the weather queries with the *basestation* infrastructure of the CDSN to investigate whether, and to what extent, users' geo-intent can help geo-map the CDSN infrastructure. For this purpose, we use a subset of basestations for which we have known GPS locations (the *ground-truth*). In order to make our analysis of (zip-code) geo-intent agnostic to diurnal and weekly variations (weekdays vs weekends), we use Dataset I exclusively in §4.1 through §4.3.

4.1 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.

Table 4 shows some of the statistics obtained using the process of correlating (zip-code) geo-intent queries with their associated basestations. We see that 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 ($\approx 23\%$). Moreover, the set formed by such basestations covers a representative fraction of SID-NID pairs, and consequently SIDs, in the network. Therefore, explicit geo-intent is pervasive not only in terms of geographic coverage (as seen in §3.2) but also in the CDSN infrastructure. This is particularly important because if geo-intent of users indeed captures their geo-location, we can possibly geo-map a significant fraction of the basestation infrastructure across wide geographies.

Fig. 3(a) and (b), respectively show the distributions for the number of user queries containing zip-codes *per basestation* and *unique* zip-codes per basestation for the URLs in Dataset I. We observe that the 50^{th} and 75^{th} percentiles for the number of queries (containing zip-codes) seen per basestation are 16 and 50 respectively. While a majority of basestations see a sizable number of geo-intent queries, the 50^{th} and 75^{th} percentiles for the number of *unique* zip-codes seen per basestation are 3 and 6 respectively. Fig. 3(c) shows the distribution of the ratio of unique zip-codes over the total number of zip-code geo-intent queries per basestation. Once again we observe that the respective 50^{th} and 75^{th} percentiles for the ratio are 0.2 and 0.4 respectively. This result clearly indicates that when there are a number of zip-code containing weather queries seen at a basestation, many of them are associated with only a small number of zip-codes. This observation has important implications in the process of geo-mapping of the basestation infrastructure, as will be explored in the next subsection.

Further analysis shows that the spread of zip-code containing geo-intent queries across the basestation infrastructure is somewhat uneven, where basestations within urban metropolitan areas generally account for a greater fraction of geo-intent queries than those in rural areas. This can partly be explained by the difference in population densities as well as the percentages of "smart" phones and data service plans adopted by users in these areas. Due to space limitation, we do not provide detailed results (area-wise statistics) here.

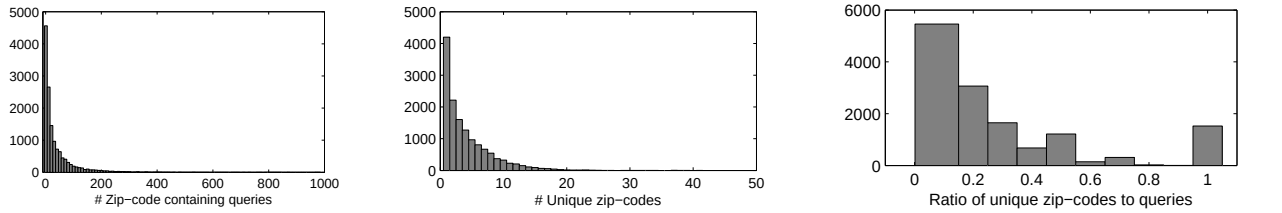
4.2 From Geo-intent to Geo-location

With about 23% of the basestations in our dataset seeing zip-code containing weather queries, can we use

[6]For example, careful analysis of the service provider, `www.geoterrestrial.com` and the queries submitted to this service reveals that running on GPS-enabled mobile devices, this service is associated with an application called *GeoToday* which provides topographical (e.g. altitude) and weather related information at the user' current location. Hence the GPS coordinates contained in user queries to this service reflect *explicit user geo-intent*, in this case, the source (user) location of the geo-intent. Similar analysis to a couple of other services also confirm that the GPS coordinates contained in the user queries also reflect true geo-intent.

Table 3: GPS coordinates in HTTP responses from web-host.

Type	Geo-physical identifiers in URL	Zoom-level
Req.	zip=53108&city=Caledonia&state=Wi&country_code=US	-
Resp.	mzip=53108&mcity=Caledonia&mstate=Wi &mx=–87.93&my=42.82	2
Resp.	mzip=53108&mcity=Caledonia&mstate=Wi & mx=–99.76 &my=42.82	1



(a) Number of zip-code queries (b) Unique zip-codes (c) Ratio of unique zip-codes to queries

Figure 3: Spread of geo-intent per basestation, Y-axis: # Basestations.

Table 4: Infrastructure coverage of zip-code geo-intent in Dataset I.

# of	Users	BSID	SID-NID	SID
Overall	2 M	62, 534	506	237
Geo-Int. (Zip)	29 K	14, 224	356	219

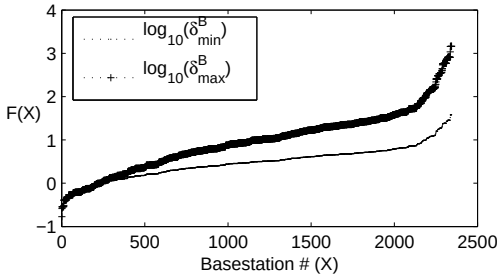


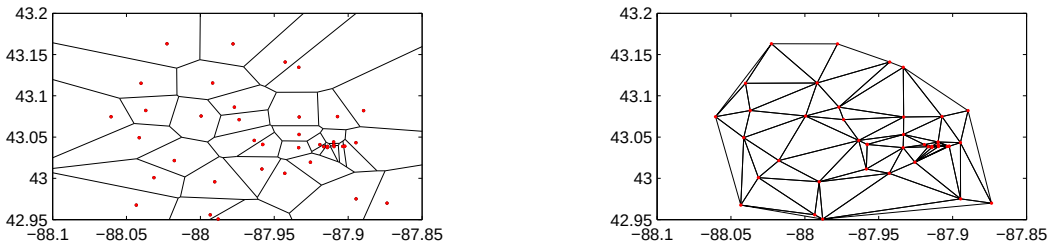
Figure 4: Distribution of δ^B_{min} and δ^B_{max} for basestations in *ground-truth-location-&-zip-code* set.

the explicit geo-intent information contained therein to geo-localize the basestations in question? We note that the zip-codes contained in users' weather queries are most likely associated with the *target* regions of users' interest; on the other hand, the basestations seeing the queries are associated with the location of users at the moment of querying. Hence the extent and accuracy of using user (explicit) geo-intent to help geo-localize the basestations will depend on how far the target location of users' interest (as specified by the zip-codes) are from the basestations where the queries are issued (the source location of users). To investigate this question, we utilize the subset of basestations for which we have the ground-truth (i.e. their GPS co-ordinates). Among the basestations with ground-truth GPS locations, we find that roughly 20% (in a similar percentage as zip-code seeing basestations to the entire basestation set) 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 ($\approx 2,400$ in all), with both the ground-truth GPS locations and associated zip-code queries, as the *ground-truth-location-&-zip-code* BSID dataset.

To examine the relationship between the locations of the basestations and users' geo-intent (the zip-codes associated with the basestations), we compute the geo-physical distances between basestations and the zip-codes as follows. 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 queried by the users associated with B. Recall that we identify each zip-code Z_i with a pair of GPS-co-ordinates for its centroid in the form of $C_{Z_i} = (lat_i, long_i)$. We denote the distance between the basestation B and the zip-code Z_i by $\delta^B_i = dist(L_B, C_{Z_i})$, where the distance is computed over the surface of the earth using the (angular) latitude and longitude co-ordinates and is then mapped to the metric distance in kilometers (km)[7]. In particular, we define $\delta^B_{min} = min_{1\leq i\leq k}\delta^B_i$ and $\delta^B_{max} = max_{1\leq i\leq k}\delta^B_i$. Further, for basestations that are associated with multiple zip-codes ($k \geq 3$), we also compute the median of δ^B_i's, denoted by δ^B_{med}. In addition, we compute the distance between each basestation and the most frequently[8] queried zip-code associated with it, and denote this distance as δ^B_*. The distributions for δ^B_{min} and δ^B_{max} are shown in Fig. 4 using the *ground-truth-location-&-zip-code* BSID dataset. For 50% of the basestations in the *ground-truth-location-&-zip-code* set, the

[7]We use the avg. value computed through the haversine and Vincenty formulas and assume a mean radius of 6,371 km for the earth.

[8]If there are two or more such zip-codes, we randomly pick one of them.



(a) Voronoi partitioning

(b) Delaunay (sub) graph

Figure 5: Milwaukee city zip-code centroids (red-dots)

distance between basestation B and the closest queried zip-code (δ^B_{min}) is within 3.5 km range, and for 75% of them it is less within 5.5 km range. In particular, about 25% of basestations lie within 1 km range of the nearest zip-code they see. This promises possibly high accuracy of geo-localizing a basestation on the basis of geo-intent in some cases. In contrast, δ^B_{max} (distance between basestation B and the farthest zip-code it sees), is within 12.5 km for 50% of basestations and within 20 km for 75% of them; much larger than corresponding δ^B_{min}. Similarly, δ^B_{med} is within 3.8 km range for 50% and within 6.9 km range for 75% of basestations while δ^B_* is within 3.9 km range for 50% and within 6.1 km range for 75% of the basestations. In short, we see that while the distance between the true location of a basestation and the farthest queried zip-code seen by it can be in the range of 10's km or more, that between the basestation and the closest queried zip-code is usually within 10 km (and often within 5 km or less). Moreover, when multiple zip-codes are queried by the users, more of them tend to be in the vicinity (around 7 km or less) of the basestation. The frequently queried zip-codes are often also the closest zip-code or a zip-code not much farther away.

However, using the absolute distance (in km) to correlate geo-intent (zip-codes) and geo-location (of basestations) does not paint the full picture, as zip-code areas have varying sizes, depending upon population density and other factors. For instance, large metropolitan cities tend to have more zip-code areas with smaller geo-physical sizes, while rural areas have far fewer, and larger, zip-code areas. To better understand the relationship between the location of basestation and the zip-codes their users query, we use the centroid of each zip-code and perform a *Voronoi partition* of the entire US mainland [9]. In other words, the US mainland is represented as a contiguous collection of Voronoi cells, where each zip-code centroid is exclusively contained in a single Voronoi cell. From the Voronoi diagram representation of the US mainland, we construct the corresponding *Delaunay graph*, in which the vertices are the zip-codes, and an edge is introduced between two zip-codes if and only if they are contained in neighboring Voronoi cells. As an example, Fig. 5 shows a portion of the Voronoi diagram (for the south-east part of Wisconsin around the Milwaukee metropolitan area) and the corresponding Delaunay (sub) graph.

We now introduce a new metric to measure the distances between basestations and zip-codes, in terms of the Voronoi diagram and the Delaunay graph introduced above, to better gauge the relationship between user geo-intent and the geo-location of the associated basestation. Given a basestation B with known GPS location $L_B = (lat_B, long_B)$, we first determine the Voronoi cell in which it lies. We associate basestation B with the zip-code contained in the same Voronoi cell, say $\widehat{Z}_B$, and refer to this zip-code as the *home* zip-code of B. Now for each zip-code Z_i seen at basestation B, we compute the (hop-count) distance between B and Z_i as the shortest path distance (in terms of hop-counts) between $\widehat{Z}_B$ and Z_i in the Delaunay graph. We denote this hop-count distance by h^B_i.

In order to understand the distribution of hop-count distances (h^B_i) between a basestation B and the zip-codes $Z_i \in Z_B$, we define a multi-hop ($l = 1, 2, 3, ...$) neighborhood relationship between the nodes of the Delaunay graph shown below:

$$Neighbor_l(\widehat{Z}_B, Z_i) = 1 \text{ if } h^B_i \leq l \quad (1)$$

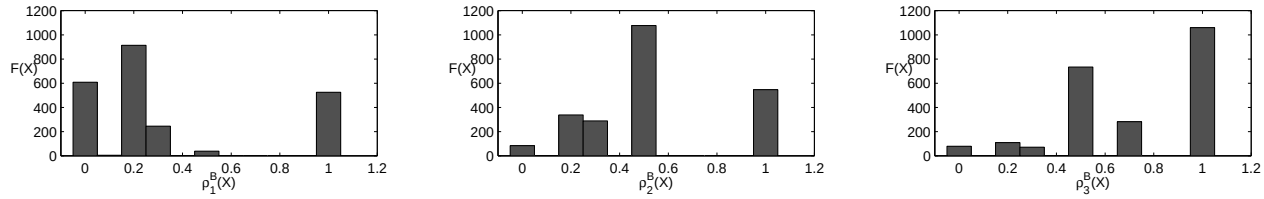
$$= 0 \text{ otherwise.} \quad (2)$$

Then, the following ratio:

$$\rho^B_l = \frac{\sum_{i=1}^{k} Neighbor_l(\widehat{Z}_B, Z_i)}{k} \quad (3)$$

where k is the number of zip-codes seen at B, provides similar insight into the hop-count distance between the home zip-code of B and the zip-codes seen at it, as the δ^B functions defined for distances over the surface of the earth. For example, ρ^B_1 tells us the fraction of zip-codes seen at B that are at most 1 hop away from

[9]Instead of partitioning the US mainland in terms of the ZCTA boundaries using data from the US census site *http://www.census.gov*, we use the Voronoi partition for ease of analysis and computation.



(a) 1 hop (b) 2 hops (c) 3 hops

Figure 6: No. of basestations (Y-axis) with X (fraction) of assoc. zip-codes at most l hops away.

B in the Delaunay graph. Fig. 6 shows the distribution of ρ_l's for $l = 1, 2, 3$. For example, in fig. 6, for a given l, the value on the Y-axis (length of the bar) corresponding to bin $X = 0.5$ on the X-axis represents the number of basestations which have 50% of the associated zip-codes at most l hops away. Notice the consistent increase in the Y-values corresponding to larger X-values as we go from $l = 1$ to $l = 3$. In fact, over 90% of the basestations have more than 50% of their associated zip-codes within $l = 3$ hops. We, therefore, (in conjunction with the evidence from similar results for absolute distances) 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.

4.3 Geo-intent, Geo-location and User Behavior

In this section we analyze user behavior, particularly in terms of user mobility, to gain further insight into the observations obtained in the preceding sections. This analysis provides a plausible explanation as to why zip-codes in weather queries – despite being associated with the *target* of geo-intent – can help geo-map the basestation infrastructure to a large extent with a reasonable accuracy, namely, within the range of 3 km to 10 km or 1-3 neighboring zip-code areas for a large majority of basestations.

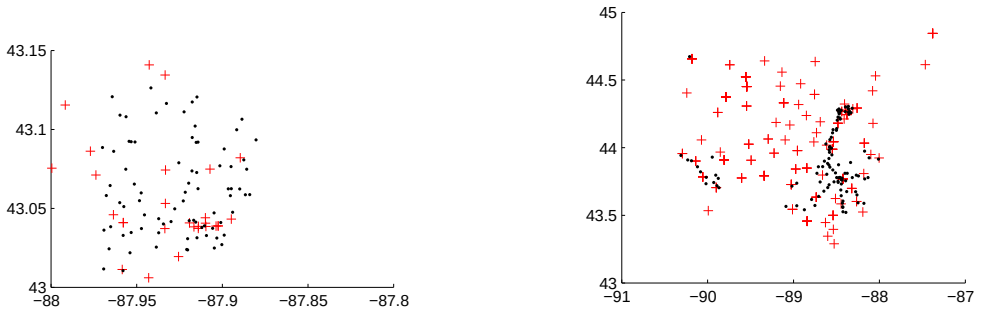
To study the user mobility behavior, we examine the number of basestations accessed by those users who express their explicit geo-intent (i.e. issuing a weather query containing a zip-code) at least once during the observation period. Using Dataset I which spans a week long period, we observe that almost 50% of the users are associated with exactly one basestation for the entire duration, while 75% of the users communicate with 4 or fewer basestations[10]. Among those who are associated with multiple (but ≤ 4) basestations, we find from the ground-truth set (when available) that such basestations are generally not far apart. This is particularly true for users within a metropolitan area. As an illustrative example, Fig. 7(a) shows a metropolitan area in the Midwest, where each black "·" indicates the location of a basestation within the metropolitan area, and each red "+" indicates the centroid of the most frequently queried zip-code by users associated with this (sub) set of basestations. We see that for a number of basestations (≈ 80), the most frequently queried zip-codes are very few (≈ 20) and confined to a small geo-physical region. In contrast, we find that basestations located in places with high user mobility, e.g. along major interstate highways, frequently see relatively greater numbers of zip-code queries from different users, but such zip-codes seem to be far more geographically dispersed (see Fig. 7(b) for an example).

Besides user mobility patterns, we also examine user query patterns. We find that among the users who express their explicit geo-intent at least once during the observation period, 90% of them query just one zip-code, while 96% query two or fewer zip-codes. In summary, our analysis shows that a majority of users tend to have limited mobility (when they access mobile data services) with respect to the basestation infrastructure (especially in metropolitan/urban areas), and their data access patterns are fairly stable with respect to the data access points (basestations). As a consequence of such user behavior and mobility patterns, namely, a majority of users tend to move around one or a few spots within a relatively limited radius and typically query for a default zip-code (close to their places of residence or work), their explicit geo-intent (in this case, the target location of their interest) can indeed help geo-localize the basestations they are associated with, albeit the extent and accuracy of the geo mapping hinges on the type of the geo-intent information available.

5. GEO-MAPPING THE BASESTATION INFRASTRUCTURE

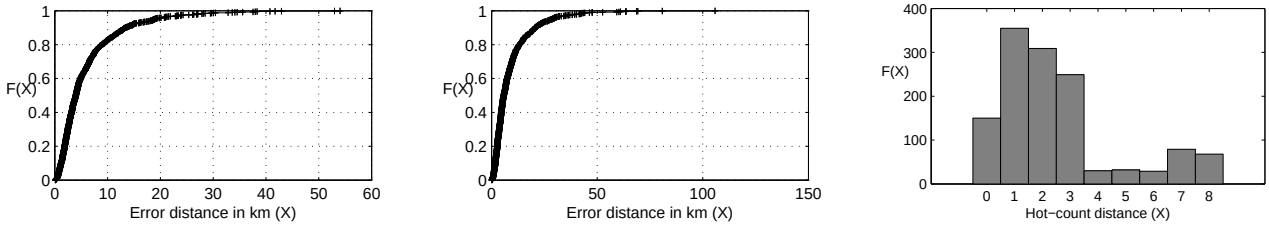
Based on the analysis and observations made in the previous section, we now present some heuristics to geo-map the basestation infrastructure. In §5.1 we first describe two simple heuristics for geo-localize the basestations which see at least one user zip-code weather query. We then extend the heuristics to geo-map those basestations that do *not* see any explicit geo-intent queries

[10]For comparison, we also perform similar analysis for those users who do *not* issue any geo-intent queries, and obtain similar results.



(a) Metropolitan area (b) Highway area

Figure 7: Most frequently queried zip-codes (red "+") seen at (sub) set of basestations (black ".").



(a) Dist. Err. (MV-scheme) (b) Dist. Err. (DNC-scheme) (c) Hop-count Err. (DNC-scheme)

Figure 8: Error incurred in direct geo-mapping compared to the ground-truth set.

but share the user base with those that do, by exploiting user movement over short time intervals. The evaluation results are presented in § 5.2. As a proof of concept, we will also present some results obtained from similar analysis applied to the GPS co-ordinates related to a particular service (`www.geoterrestrial.com`) from Dataset II in §5.3.

5.1 Geo-Mapping Heuristics

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 the user explicit geo-intent by means of the following two simple 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 either equally or proportionally among all the zip-codes s/he queries. For example, if U_i queries zip-code Z_1 thrice and Z_2 twice, in equal vote-splitting, both zip-codes receive 0.5 votes from U_i while in proportional vote-splitting, 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 the basestation B. When there are multiple winners (ties), all of them are chosen as probable locations (with equal probability).

The second heuristic, the *Dense Neighborhood Clustering* (DNC) scheme, uses both the frequencies of zip-codes queried as well as the neighborhood relationship among the zip-codes. Given the Delaunay graph of the US mainland, the zip-codes in Z_B induce a subgraph, denoted by $G_Z(B)$, with vertices Z_i, $1 \leq i \leq k$, and there is an edge between Z_i and Z_j if and only if the zip-code areas they represent border each other. Furthermore, we assign each node Z_i a weight w_i equal to the votes received by it during the Majority Voting scheme above. In general, the subgraph $G_Z(B)$ consists of multiple connected components, $C_1, \ldots, C_m$, where $1 \leq m \leq k$ ($k = |Z_B|$), each a probable candidate for the location of B. For each C_p, we define $w(C_p) = \sum_{Z_i \in C_p} w(Z_i)$. We select the component C_p with the largest $w(C_p)$ as the probable location (a connected zip-code neighborhood) for the basestation B. We note that in the special case where $m = k$, i.e. $G_Z(B)$ consists of k disjoint vertices, this scheme reduces to Majority Voting. A further refinement of this heuristic also filters out cases where the total weight $w(C_p)$ of the winner component is too small (below a threshold) and $G_Z(B)$ consists of mostly disconnected

vertices that are spread over a large geographical area. In such cases, the heuristic simply labels the location of B as "undecided" instead.

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[11] 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 two direct geo-mapping heuristics 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 $h(B, \mathcal{B}_{mapped}) = \min_{\overline{B} \in \mathcal{B}_{mapped}} h(B, \overline{B})$. 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. To control the mapping accuracy of a basestation $B \in \mathcal{B}_{unmapped}$, we introduce two parameters, the *hop-count threshold* d, and the *(mapped) neighborhood size* s. For any $B \in \mathcal{B}_{unmapped}$ such that $h(B, \mathcal{B}_{mapped}) \leq d$ and 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. Alternatively, we construct a connected zip-code neighborhood (subgraph), also denoted by H_B, which is formed by the zip-codes whose centroids fall within the convex hull H_B. We refer to H_B as the inferred zip-code neighborhood for basestation B.

[11] More generally, for each edge $e = (B_i, B_j)$, we count the number of users associated with both B_1 and B_2 within a short time interval ΔT, and filter out edges that have a very small common user count to prevent spurious connections due to noisy data. If a lot of users access B_i and B_j within a short interval they are likely close to each other

5.2 Evaluation and Validation

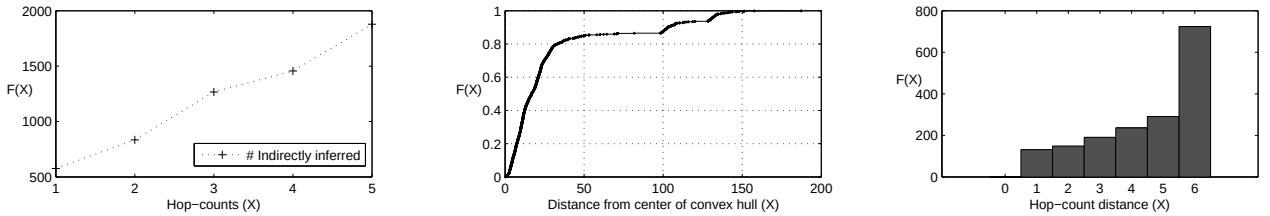
To evaluate the efficacy of our heuristics for geo-mapping the basestation infrastructure, we use the collection of basestations with *ground-truth* GPS locations. In particular, we use the basestations in the *ground-truth-&-zip-code* set for Dataset II to evaluate the two direct geo-mapping heuristics. Then, we use the other basestations in the *ground truth* set, which do not see zip-code queries by their users, in both Dataset I and Dataset II to evaluate the indirect mapping heuristics.

Using the *ground-truth-&-zip-code* basestation dataset from Dataset II, Fig. 8(a) shows the distribution of geo-mapping errors, namely, the distance between the inferred location and the ground-truth location, using the *Majority Voting* heuristic. In case of multiple inferred locations (zip-codes) available for a basestation, we compute the error for each inferred location. We observe that the 50^{th} and 75^{th} percentiles are around 3.9 and 6.1 km respectively (Dataset II), quite similar to what we observed for δ^B_{min} in Dataset I (see § 4.2).

For the *Dense Neighborhood Clustering* heuristic, we measure the errors in terms of both absolute distance and hop-count. For a basestation B, let $C(B)$ be the inferred zip-code neighborhood. We compute the centroid of $C(B)$ and use the distance (ground-truth GPS location) between B and the centroid as the absolute distance error. In terms of hop-count distance error, we use the home zip-code $\widehat{Z}_B$ of the basestation B, and compute the (shortest distance) hop-count from $\widehat{Z}_B$ to $C(B)$, namely, $h(\widehat{Z}_B, C_B)$ as defined in the indirect geo-mapping heuristics. Figs. 8(b) and (c) respectively show the distributions of absolute and hop-count errors. We see that the errors in absolute distances are comparable to those obtained for absolute distances in Majority Voting for most basestations. This is not surprising as most of the clusters in our dataset are of small sizes (made of 4 or less zip-codes) and span relatively small geo-graphical areas, especially in urban locations. We also see that the 50^{th} and 75^{th} percentiles for the hop-count distance are 2 and 3 respectively.

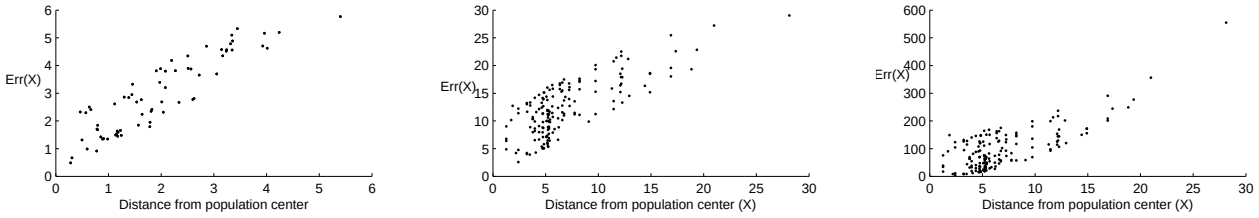
Next, we evaluate the heuristic for the indirect geo-mapping of basestations in $\mathcal{B}_{unmapped}$. To do so, we fix the (mapped) neighborhood size s to 3[12], and vary the hop-count threshold d from 1 to 5. We measure the geo-mapping errors in terms of the absolute distance (i.e. the distance from the ground-truth GPS location of B to the centroid of the inferred convex hull H_B) and hop-counts (i.e. $h(\widehat{Z}_B, H_B)$). Fig. 9(a) shows the percentage of additional basestations that can be indirectly geo-mapped as a function of d. Fig. 9(b) and (c)

[12] In our dataset, increasing s does not significantly improve the accuracy of the geo-mapping while considerably reduces the coverage



(a) No. of unmapped BSID's covered (b) Dist. Err. (DNC-scheme) (c) Hop-count Err. (DNC-scheme)

Figure 9: Coverage and error incurred in indirect geo-mapping compared to the ground-truth set.



(a) Metropolitan area (b) Urban area (c) Highway

Figure 10: Relationship between population and basestation density and error in geo-mapping.

show the distributions of absolute distance errors and hop-count errors, respectively. We see that the errors incurred go up in terms of distances, even though hop-counts go up only by a few hops. The reason for this is that the edges in user mobility graph cover long distances in highway areas. Add to it, the long distances between the vertices of the convex hull H_B in such areas due to far apart zip-codes. Even for small values of d and s, the error incurred in such areas is high. In contrast, in urban areas, users cannot travel very long distances in short intervals of time. This means shorter edges in the mobility graph. Also, zip-codes in such areas are geo-physically close to each other. A combination of the two results in relatively lower error in urban areas for the indirect method both in terms of hop-counts (usually $1-3$) as well as absolute error ($5-10$ km or less). This helps us realize our objective of geo-localizing basestations in $\mathcal{B}_{unmapped}$ at city level granularities.

To explore the effects of population density in a region, which determines both the size of zip-codes and basestation densities, on the accuracy of geo-mapping, we conduct a case study. We select three non-overlapping regions - a metropolitan area (34 zip-codes, 235 basestations), an urban area with a relatively lower population density than a metropolitan area (20 zip-codes, 115 basestations), and a stretch of an interstate highway connecting several urban centers along south-eastern Wisconsin (6 towns, 130 basestations). In each case, we identify high population density centers (7 most populous zip-codes in the metro, 5 in the urban area and the centroids of the 6 towns in the case of the highway). Fig. 10 shows the (absolute distance) errors incurred in geo-mapping basestations via explicit geo-intent in all the three cases as a function of distance between the basestations and the nearest population center. We see that the error varies almost linearly with the distance from the population centers in the case of metropolitan and the urban areas (average errors smaller for the metropolitan area than the urban area). This seems to show that users usually query for information in and around themselves for some preferred target locations (e.g. downtown, residential areas) represented by the population centers. On the other hand, the errors incurred in the case of the highway are substantially high even for relatively short distances from the nearest city in the vicinity. This is possibly because people usually query information related to the regions they are coming from or, more likely, going towards, while on the highway.

5.3 Geo-mapping using GPS Geo-intent

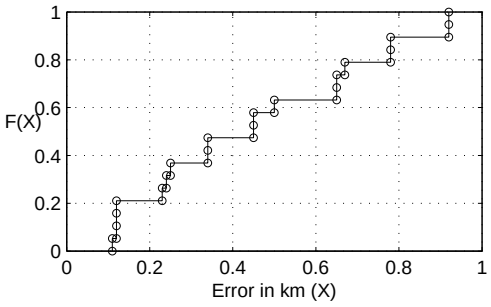
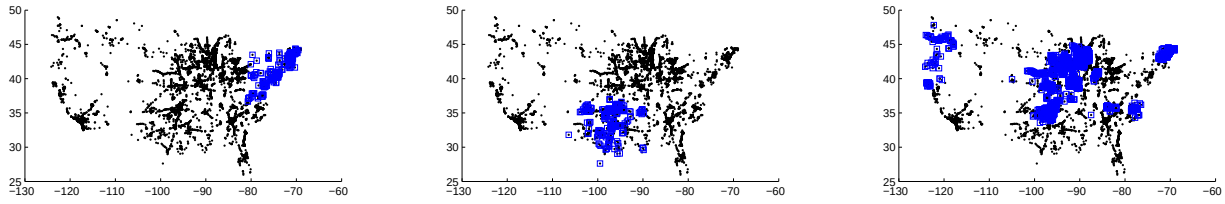


Figure 11: CDF of the mean errors (km) in geo-mapping using the small GPS geo-intent data.



(a) (b) (c)

Figure 12: Geo-physical coverage of NAS gateways.

Lastly, we use the small *GPS geo-intent* dataset discussed in section §3.3 to illustrate the efficacy of our approach when GPS-based geo-intent (in particular, when the GPS coordinates are associated with the source (user) locations of the geo-intent. We extend the direct geo-mapping heuristics from §5.1 to the case of GPS coordinates, 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. 11 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.

6. IP INFRASTRUCTURE IN THE CDSN

In this section we examine the IP network infrastructure in the CDSN with the help of inferred (and ground-truth) locations for the set of geo-mapped basestations. Our goal is two-fold: i) to infer and understand the relationship (e.g. the geo-spatial distribution) of the IP network elements with the basestation substrate, and ii) to investigate whether we can geo-map the network elements in the CDSN to some degree. Here, we use the packet data sessions to extract the relevant relationships. Recall, the RADIUS/RADA packets contain basestation ids. In addition, they also contain four types of IP addresses of interest to us[13]. They are: *framed* IP addresses (assigned to end users' devices); the RADIUS/RADA server IP addresses (assigned to servers responsible for authenticating a user's session), IP addresses of *NAS gateways* (gateway servers to the IP data distribution backbone network in the CSDN) and IP addresses of home-agents (servers that maintain certain user information e.g. user registration, credentials, and current locations for mobile IP routing).

Not surprisingly, a predominant majority of the IP addresses in the datasets are framed IP addresses. They mostly come from the private address realm of the IP space; this is consistent with the findings in [3] where the authors collect and analyze the IP addresses seen at the end users' devices. The framed IP addresses appear to be assigned randomly from the private address ranges agnostic to the geo-location of the basestations. The number of the other three IP addresses are far smaller: both NAS gateways and home-agent IP addresses number within 100, and RADIUS/RADA servers below 10 – in stark contrast to the number of basestations (in tens of thousands). As the NAS gateways and home-agents are more likely to correlate with user/basestation locations, in the following we explore the geo-spatial distribution of these IP addresses and their relation to the basestation infrastructure.

For each NAS gateway/home-agent IP address, we extract all the BSIDs which appear in the same RADIUS/RADA data packets containing the said IP address – these basestations are where the user data sessions originate. Hence each IP address (NAS gateway server or home-agent) is associated with a collection of basestations. We study the geo-spatial distribution of these basestations to investigate whether there is any significant geo-spatial correlation between the NAS gateways/home agents and locations of the basestations. We further analyze the relationships between NAS gateways by clustering them based on the number of basestations they share in common, i.e. the size of intersection between the basestation collections associated with the two NAS gateway IP addresses.

As representative examples, Fig. 12 depicts the geo-spatial distribution of the basestations associated with three different NAS gateways. We observe that these NAS gateways cover rather large geographical areas (spanning multiple states, and in terms of the basestation infrastructure, multiple SIDs). These areas are typically contiguous (as in Figs. 12 (a) and (b)), but sometimes can be disparate too (as in Fig. 12(c)). Furthermore, two or more NAS gateways may share a large

[13]Note that since the data is collected through *passive* measurement, we do not have IP addresses of IP routers, DNS servers, etc., and we are unable to conduct active measurements in the CDSN.

overlapping set of basestations; it appears that these gateways serve the same large geographical region for load-balancing. We also performed similar analysis for home-agent IP addresses (where we also take into account the user activities to account for user mobility and roaming). We find that each home-agent IP address also covers a large geographical region, and multiple home-agents may cover the same or similar regions for load-balancing. Due to space limitation, we do not provide these results here.

In summary, we find that in contrast to the basestation infrastructure, the numbers of NAS gateways and home-agents are far smaller. While these gateways/home-agents are geo-spatially distributed, each covers a large geographical region spanning multiple states and corresponds to a large collection of the basestations in the CDSN substrate. Our findings point to several challenges in attempting to geo-map the CDSN infrastructure *from the outside* (cf. [3]), and in deploying *location-aware* content distribution services *outside the CDSN* to serve users inside the CDSN.

Last but not the least, we remark that comparing the two datasets collected about 10 months apart, we observe that the amount of cellular data activity and traffic has grown tremendously: for instance, the numbers of data sessions and application sessions increased over 3 and 10 times, respectively (see table 1). Moreover, the number of cellular data users have also increased considerably. With the increasing popularity of new generations of smart phones, the growth in cellular data traffic will further spur expansion of the IP networks within cellular service provider networks, and we may see more finer-grain geographic coverage to better cater to the growing user demand within a CSDN infrastructure.

7. CONCLUSION AND FUTURE WORK

In this paper we put forth a novel approach for mapping the CDSN infrastructure via *(explicit) user geo-intent*, which circumvents the challenges plaguing conventional approaches (e.g. active probing). 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 and gateway IP addresses to geo-map the CDSN infrastructure. To investigate whether — and to what extent — our approach can help geo-map the CDSN infrastructure, we employed the data (RADIUS/RADA packet data sessions and HTTP application sessions) collected at the core IP network inside a CDSN. We developed heuristics for identifying user geo-intent to geo-map the CDSN infrastructure — in particular, the basestations and IP NAS gateways — 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 $3.9 - 6.1$ km range in general and within $1.5 - 2$ km range in densely populated urban areas. Furthermore, the geo-mapping accuracy is far better (often within 1 km) in large metro-areas with dense population and smaller zip-code areas. Using the inferred and ground-truth GSP locations of the basestations, we also examined the geo-spatial distribution of IP network elements such as NAS gateways and IP home-agents, and their relationship to the cellular network substrate.

With growing popularity of newer generations of GPS-enabled smart phones and increasing prevalence of location-specific and location-aware services and apps, we expect our geo-intent based mapping approach to yield more precise results, as was illustrated using a small set of user geo-intent queries with GPS coordinates in our datasets.

Given the unprecedented growth in cellular data traffic, 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 in this direction and much additional research is still sorely needed.

8. REFERENCES

[1] *http://www.census.gov.*

[2] *http://www.roamingzone.com.*

[3] M. Balakrishnan, I. Mohomed, and V. Ramasubramanian, *Where's that phone? Geolocating IP addresses on 3g networks*, Proc. of ACM Internet Measurement Conference (2009).

[4] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, *Analysis of geographic queries in a search engine log*, Proc. of LocWeb (2008).

[5] C. Rigney, *RADIUS accounting*, Internet RFC 2866 (2000).

[6] C. Rigney, S. Willens, A. Rubens, and W. Simpson, *Remote authentication dial in user service (RADIUS)*, Internet RFC 2865 (2000).

[7] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, *Measuring serendipity: Connecting people, locations and interest in a mobile 3G network*, Proc. of ACM Internet Measurement Conference (2009).

[8] X. Yi, H. Raghavan, and C. Leggetter, *Discovering user's specific geo intention in web search*, Proc. of World Wide Web (2009).

[9] H. Zang, F. Baccelli, and J. C. Bolot, *Bayesian inference for localization in cellular networks*, Proc. of IEEE INFOCOM 2010, March 2010.