Cellular Data Network Infrastructure Characterization and Implication on Mobile Content Placement

Qiang Xu
University of Michigan
qiangxu@eecs.umich.edu

Junxian Huang
University of Michigan
hjx@eecs.umich.edu

Zhaoguang Wang
University of Michigan
zgw@eecs.umich.edu

Feng Qian
University of Michigan
fengqian@eecs.umich.edu

Alexandre Gerber
AT&T Labs Research
gerber@research.att.com

Z. Morley Mao
University of Michigan
zmao@eecs.umich.edu

ABSTRACT

Despite the tremendous growth in the cellular data network usage due to the popularity of smartphones, so far there is rather limited understanding of the network infrastructure of various cellular carriers. Understanding the infrastructure characteristics such as the network topology, routing design, address allocation, and DNS service configuration is essential for predicting, diagnosing, and improving cellular network services, as well as for delivering content to the growing population of mobile wireless users. In this work, we propose a novel approach for discovering cellular infrastructure by intelligently combining several data sources, *i.e.,* server logs from a popular location search application, active measurements results collected from smartphone users, DNS request logs from a DNS authoritative server, and publicly available routing updates. We perform the first comprehensive analysis to characterize the cellular data network infrastructure of four major cellular carriers within the U.S. in our study.

We conclude among other previously little known results that the current routing of cellular data traffic is quite restricted, as it must traverse a rather limited number (*i.e.,* 4–6) of infrastructure locations (*i.e.,* GGSNs), which is in sharp contrast to wireline Internet traffic. We demonstrate how such findings have direct implications on important decisions such as mobile content placement and content server selection. We observe that although the local DNS server is a coarse-grained approximation on the user's network location, for some carriers, choosing content servers based on the local DNS server is accurate enough due to the restricted routing in cellular networks. Placing content servers close to GGSNs can potentially reduce the end-to-end latency by more than 50% excluding the variability from air interface.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless communication; C.2.3 [**Network Operations**]: Network monitoring; C.4 [**Performance of Systems**]: Measurement techniques; C.4 [**Performance of Systems**]: Reliability, availability, and serviceability

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.

General Terms

Experimentation, Measurement, Performance

Keywords

Cellular network architecture, GGSN placement, Mobile content delivery

1. INTRODUCTION

On the Internet, IP addresses indicate to some degree the identity and location of end-hosts. IP-based geolocation is widely used in different types of network applications such as content customization and server selection. Using IP addresses to geolocate wireline end-hosts is known to work reasonably well despite the prevalence of NAT, since most NAT boxes consist of only a few hosts [7]. However, one recent study [5] exposed very different characteristics of IP addresses in cellular networks, *i.e.,* cellular IP addresses can be shared across geographically very disjoint regions within a short time duration. This observation suggests that cellular IP addresses do not contain enough geographic information at a sufficiently high fidelity. Moreover, it implies only a few IP gateways may exist for cellular data networks, and that IP address management is much more centralized than that for wireline networks, for which tens to hundreds of Points of Presence (PoPs) are spread out at geographically distinct locations.

There is a growing need to improve mobile content delivery, *e.g.,* via a content distribution network (CDN) service, given the rapidly increasing mobile traffic volume and the fact that the performance perceived by mobile users is still much worse than that for DSL/Cable wireline services [17]. For mobile content, the radio access network, cellular backbone, and the Internet wireline all have impact and leave space for further improvement [2, 1, 30]. A first necessary step is to understand the cellular network structure.

The lack of geographic information of cellular IP addresses brings new challenges for mobile service providers, who attempt to deliver content from servers close to users. First, it is unclear where to place the content servers. As shown later, cellular data networks have very few IP gateways. Therefore, it is critical to first identify those IP gateways to help decide where to place content servers. Second, unlike wireline networks, cellular IP addresses themselves often cannot accurately convey a user's location, which is critical information needed by the CDN service to determine the closest server. In this work, we show how these challenges can be addressed by leveraging the knowledge of the cellular network infrastructure.

Cellular data networks have not been explored much by the re-

search community to explain the dynamics of cellular IP addresses despite the growing popularity of their use. The impact of the cellular architecture on the performance of a diverse set of smartphone network applications and on cellular users has been largely overlooked. In this study, we perform the first comprehensive characterization study of the cellular data network infrastructure to explain the diverse geographic distribution of cellular IP addresses, and to highlight the key importance of the design decisions of the network infrastructure that affect the performance, manageability, and evolvability of the network architecture. Understanding the current architecture of cellular data networks is critical for future improvement.

Since the observation of the diversity in the geographic distribution of cellular IP address in the previous study [5] indicates that there may exist very few cellular IP data network gateways, identifying the location of these gateways becomes the key for cellular infrastructure characterization in our study. The major challenge is exacerbated by the lack of openness of such networks. We are unable to infer topological information using existing probing tools. For example, merely sending traceroute probes from cellular devices to the Internet IP addresses exposes mostly private IP addresses along the path within the UMTS architecture. In the reverse direction, only some of the IP hops outside the cellular networks respond to traceroute probes.

To tackle these challenges, instead of relying on those cellular IP hops, we use the geographic coverage of cellular IP addresses to infer the placement of IP gateways following the intuition that those cellular IP addresses with the same geographic coverage are likely to have the same IP allocation policy, *i.e.*, they are managed by the same set of gateways. To obtain the geographic coverage, we use two distinct data sources and devise a systematic approach for processing the data reconciling potential conflicts, combined with other data obtained via simple probing and passive data analysis. Our approach of deploying a lightweight measurement tool on smartphones provides the network information from the perspective of cellular users. Combining this data source with a location search service of a cellular content provider further enhances our visibility into the cellular network infrastructure.

One key contribution of our work is the measurement methodology for characterizing the cellular network infrastructure, which requires finding the relevant address blocks, locating them, and clustering them based on their geographic coverage. This enables the identification of the IP gateways within cellular data networks, corresponding to the first several outbound IP hops used to reach the rest of the Internet. We draw parallels with many past studies in the Internet topology characterization, such as the Rocketfuel project [31] characterizing ISP topologies, while our problem highlights additional challenges due to the lack of publicly available information and the difficulties in collecting relevant measurement data. We enumerate our key findings and major contributions below.

- We designed and evaluated a general technique for distinguishing cellular users from WiFi users using smartphones and further differentiating network carriers based on cellular IP addresses. Compared with other heuristics such as querying IP addresses from *whois* database and distinguishing cellular carriers based on key words such as "mobility" and "wireless" from the organization name, our technique collects the ground truth observed by smartphone devices by deploying a lightweight measurement tool for popular smartphone OSes. Distributed as a free application on major smartphone application markets, it can tell the carrier name for 99.97% records of a popular location search application which has 20,000 times more records than the application.

- We comprehensively characterized the cellular network infrastructure for four major U.S. carriers including both UMTS and EVDO networks by clustering their IP addresses based on their geographic coverage. Our technique relies on the device-side IP behavior easily collected through our lightweight measurement tool instead of requiring any proprietary information from network providers. Our characterization methodology is applicable to all cellular access technologies (2G, 3G, or 4G).

- We observed that the traffic for all four carriers traverses through only 4–6 IP gateways, each encompassing a large geographic coverage, implying the sharing of address blocks within the same geographic area. This is fundamentally different from wireline networks with more distributed infrastructure. The restricted routing topology for cellular networks creates new challenges for applications such as CDN service.

- We performed the first study to examine the geographic coverage of local DNS servers and discussed in depth its implication on content server selection. We observe that although local DNS servers provide coarse-grained approximation for users' network location, for some carriers, choosing content servers based on local DNS servers is reasonably accurate for the current cellular infrastructure due to restricted routing in cellular networks.

- We investigated the performance in terms of end-to-end delay for current content delivery networks and evaluated the benefit of placing content servers at different network locations, *i.e.*, on the Internet or inside cellular networks. We observed that pushing content close to GGSNs can potentially reduce the end-to-end latency by 50% excluding the variability from air interface. Our observation strongly encourages CDN service providers to place content servers inside cellular networks for better performance.

The rest of this paper is organized as follows. We first describe related work in §2. §3 describes the high-level solution to discover IP gateways in cellular infrastructure. §4 explains the main methodology in the data analysis and the data sets studied. The results in characterizing cellular data network infrastructure along the dimensions of IP address, topology, local DNS server, and routing behavior are covered in §5. We discuss the implications of these results in §6 and conclude in §7 with key observations and insights on future work.

2. RELATED WORK

Our study is motivated by numerous previous measurement studies [34, 38, 22], *e.g.*, Rocketfuel [31] to characterize various properties of the Internet through passive monitoring using data such as server logs and packet traces, as well as active measurement such as probing path changes. Efforts on reverse engineering properties of the Internet [32] have been shown to be quite successful; however, very little work has been done in the space of cellular IP networks. Complementary to our study, the most recent work by Keralapura *et al.* profiled the browsing behavior by investigating whether there exists distinct behavior pattern among mobile users [20]. Their study implemented effective co-clustering on large scale user-level web browsing traces collected from one cellular provider. As far as

we know, our study is the first to comprehensively characterize cellular IP networks covering all the major cellular carriers in the U.S., focusing on key characteristics such as network topological properties and dynamic routing behavior. From the characterization of the cellular data network structure, we also draw conclusions on content placement, which is essential given the rapidly growing demand for mobile data access.

We build our work upon a recent study by Balakrishnan *et al.* [5] in which they highlighted unexpected dynamic behavior of cellular IP addresses. Our work performs a more complete and general study covering a wider set of properties, illustrating carrier-specific network differences, explaining the observed diverse geographic distribution of cellular IP addresses, also investigating associated implications of observed network designs.

Although there have been studies characterized the CDNs relative to the end users accessing from the wireline networks [23, 29, 10, 26, 16], very little attention has been paid to the cellular users. These previous studies are mainly from two perspectives, *i.e.*, content placement and server selection. Our work is complementary to these studies by investigating the implication of cellular network infrastructure on mobile data placement and server selection. To our best knowledge, our study is the first to investigate the content placement and content server selections for cellular users.

Previous studies on cellular networks can be classified approximately into several categories, namely from ISP's view point of managing network resources [13, 33], from end-user's perspectives of optimizing energy efficiency and network performance at the device [4, 39, 6], and finally developing infrastructure support for improving mobile application performance [8, 28] and security [24]. Our work is complementary to them by exposing the internal design of the cellular data network structure that can be useful to guide such optimization efforts.

There have also been several measurement studies in understanding the performance and usage of cellular networks. One recent study focuses on mobile user behavior from the perspective of applications such as [36] which characterized the relationship between users' application interests and mobility. Other examples include a study of the interaction between the wireless channels and applications [21], performance study of multimedia streaming [11], and performance of TCP/IP over 3G wireless with rate and delay variation [9]. Note that our work fills an important void in the space of cellular data network by focusing on the network architectural design: IP address allocation, local DNS service setup, and routing dynamics.

3. OVERVIEW

In this section, we describe the cellular data network architecture, followed by an overview of our methodology for characterizing the cellular data network infrastructure.

3.1 Cellular Data Network Architecture

Despite the difference among cellular technologies, a cellular data network is usually divided into two parts, the Radio Access Network (RAN) and the Core Network. The RAN contains different infrastructures supporting 2G technologies (*e.g.*, GPRS, EDGE, 1xRTT, *etc.*) and 3G technologies (*e.g.*, UMTS, EVDO, *etc.*), but the structure of the core network does not differentiate between 2G and 3G technologies. In this study we focus on the core network, in particular, the gateways that hide the cellular infrastructures from the external network, as identifying the gateways is the key to explain the geographically diverse distribution of cellular addresses [5].

Figure 1 illustrates the typical UMTS/EDGE network. The RAN

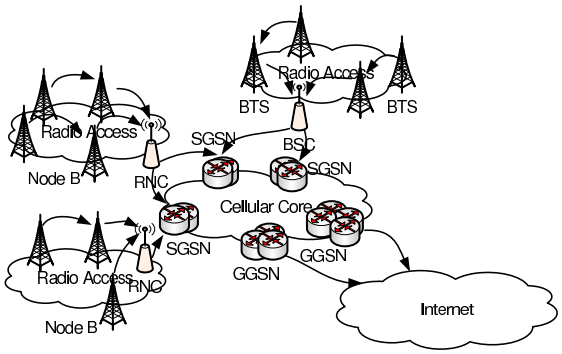


Figure 1: The UMTS/EDGE network architecture.

architecture, which allows the connectivity between user handsets and the core network, depends on the radio access technology: it consists of the Base Transceiver Station (BTS) and the Base Station Controller (BSC) for EDGE (2G), and the Node B and the Radio Network Controller (RNC) for UMTS (3G). The core network, which is shared by both 2G's and 3G's RANs, is comprised of the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN). To start a data session, a user first communicates with its local SGSN that delivers its traffic to a GGSN. The SGSN requests the DNS server for the GGSN via the user's access point name (APN). The DNS server decides which GGSN serves the data session accordingly [27]. Once the GGSN is determined, the communication between the SGSN and the GGSN is tunneled, so GGSN is the first IP hop and is followed by multiple hops such as NAT and firewalls within the core network. Being the first router for the connected cellular device, the GGSN is responsible for IP address assignment, IP pool management, address mapping, QoS, authentication, *etc.* [12].

The EVDO network has an architecture very similar to the UMTS network except that the Packet Data Serving Node (PDSN) in the EVDO core network serves as a combination of both the SGSN and the GGSN in the UMTS core network. Without explicit explanation, our statements for the UMTS network are applicable to the EVDO network as well.

To support the future 4G LTE (Long Term Evolution) network, the GGSN node will be upgraded to a common anchor point and gateway (GW) node, which also provides backward compatibility to other access technologies such as EDGE and UMTS [18]. The functionality of GW is largely similar to that of GGSN. Therefore, our proposed methodology, which focuses on identifying GGSNs in the cellular infrastructure, is still broadly applicable.

3.2 Solution Overview

Despite the growing popularity of smartphones, cellular data networks have not been explored much by the research community to explain the dynamics of cellular IP addresses. Besides the challenge of keeping tracking of cellular IP addresses, to identify GGSNs in the cellular infrastructure, another challenge is the lack of openness of such networks. Outbound probing via traceroute from the cellular devices to the Internet IP addresses exposes mostly private IP addresses along the path for UMTS networks due to the placement of NAT boxes. These NAT boxes and firewalls prevent the inbound traceroute probing to reach into the cellular backbones as well.

Identifying GGSNs in the cellular infrastructure is the key to explain the geographically diverse distribution of cellular addresses discovered by the recent study [5]. GGSNs serve as the gateway

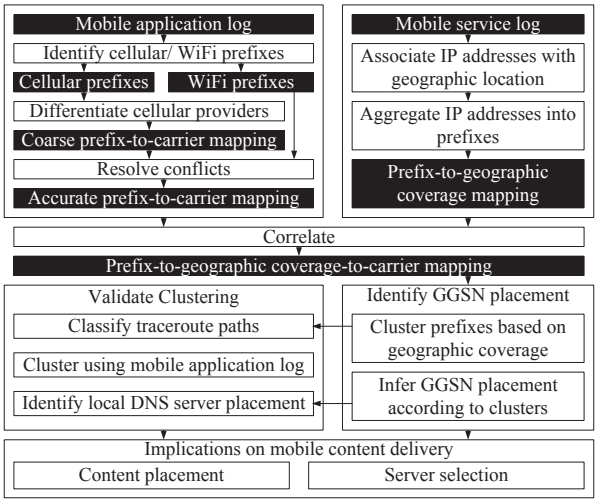


Figure 2: Workflow of the solution: black boxes correspond to data, manipulated by processing modules (white boxes).

between the cellular and the Internet infrastructure and thus play an essential role in determining the basic network functions, *e.g.*, routing and address allocation. In our study, we leverage the geographic coverage of cellular addresses to infer the placement of GGSNs, assuming that prefixes sharing similar IP behaviors are likely to have the same IP allocation policy, *i.e.*, they are managed by the same GGSN. Considering geographic coverage as one type of IP behaviors, we cluster prefixes based on the feature of geographic coverage, and infer the placement of GGSNs according to the prefix clusters that we generated.

As depicted by Figure 2, to get the geographic location of cellular IP addresses, we leverage a popular location search service whose server logs public IP address and GPS location of users (*i.e.*, the mobile service log). We use the mobile service log to generate prefix-to-geographic-coverage mappings. In order to identify cellular addresses and to further differentiate different cellular providers, we also deploy a measurement tool for mainstream smartphone OSes to build a database (*i.e.*, mobile application log) for prefix-to-carrier mappings. They provide the ground truth for determining the cellular provider who owns a certain IP address block. By correlating prefix-carrier mappings with prefix-geographic coverage mappings, we can obtain the prefix-to-geographic-coverage-to-carrier mappings for clustering. Once the clustering is finished, we validate the clustering results via three independent ways: clustering using the mobile application log, identifying the placement of local DNS servers in cellular networks, and classifying traceroute paths. Based on our findings during clustering and validation, we investigate implications of the cellular infrastructure on content delivery service for mobile users.

Note that we designed our methodology to be generally applicable for any data cellular network technologies (2G, 3G, and 4G), and particularly from the perspective of data requirement. Any mobile data source that contains IP addresses, location information, and network carrier information can be used for our purpose of characterizing the cellular data network infrastructure. Based on our experience of deploying smartphone applications, it is not difficult to collect such data.

In §4, we detail our methodology for identifying cellular addresses and cellular providers. The discovery of the geographic coverage of cellular prefixes and clustering techniques are elaborated in §5.1.

4. MEASUREMENT DATA AND PREPROCESSING

In this section, we describe the data sets used for analysis and the additional experiments carried out to supplement these data for identifying the key properties of interest. Note that due to privacy concerns, without compromising the usefulness of the results, we have anonymized the carrier by assigning a letter, *i.e.*, Carrier A through D, to identify each of the four carriers studied which have significant footprint in the U.S. Similarly, we assign a unique ID to each address block and assign a symbol to each ASN.

4.1 DataSource1 – server logs

Operator	3G	Records (%)	# BGP prefixes	# /24 prefixes	# ASNs
Carrier A	UMTS	43.34%	54	16,288	1
Carrier B	UMTS	7.09%	12	41	1
Carrier C	EVDO	1.51%	202	15,590	2
Carrier D	EVDO	1.22%	172	11,205	1
*	-	100%	16,439	121,567	1,862

Table 1: Statistics of *DataSource1*(server logs).

The first data set used is from server logs associated with a popular location search service for mobile users. We refer to this data source as *DataSource1*. It contains the IP address, the timestamp, and the GPS location of mobile devices. The GPS location is requested by the application and is measured from the device. The data set ranges from August 2009 until September 2010, containing several million records. This comprehensive data set covers 16,439 BGP prefixes, 121,567 /24 address blocks from 1,862 AS numbers. However, *DataSource1* does not differentiate the carrier for each record. Later we discuss how to map *DataSource1*'s records to corresponding cellular carriers or WiFi networks with the help of *DataSource2*'s prefix-to-carrier table in §4.3. Users of the search service may also use WiFi besides cellular networks to access the service.

Table 1 shows the breakdown of the records among the four major U.S. cellular providers for *DataSource1*. 43.34% of all the records in *DataSource1* are mapped to Carrier A due to the disproportionate popularity of the service among different mobile users. Despite this bias, we still find sufficient information to characterize the other three major carriers. 46.71% of *DataSource1* is from WiFi users, and 0.13% is from cellular carriers besides the four major carriers. Note that one cellular carrier may be mapped to more than one AS number (ASN), *e.g.*, Carrier C corresponds to more than one ASN.

The long-term and nation-wide *DataSource1* is the major data source that we rely on to map cellular prefixes to their geographic coverage after we aggregate cellular IP addresses to prefixes based on *RouteViews*'s BGP update announcements [3].

4.2 DataSource2 – active measurements

The second main data source of our analysis comes from an application that we have widely deployed on three popular smartphone platforms: iPhone OS, Android, and Windows Mobile (WM). We refer to this data source as *DataSource2*, with the basic statistics shown in Table 2. The application is freely available for mobile users to download for the purpose of evaluating and diagnosing their networks from which we can collect common network char-

Platform	# users	# carriers	# BGP prefixes	# /24 prefixes	# ASNs
iPhone	25K	- [1]	5.2(1.8)K	10.8(2.8)K	1.2K(268)
Android	28K	278(36)[2]	2.7(1.1)K	7.3(3.1)K	720(179)
WM	9K	516(66)	1.6(0.5)K	5.7(3.5)K	545(121)
other	63K	571(87)	7.6(2.9)K	23(9.3)K	1.5K(387)

[1] On iPhone OS, we cannot tell the serving carrier.
[2] Numbers inside parentheses refer to the U.S. users only.

Table 2: Statistics of *DataSource2*(smartphone app).

acteristics such as the IP address, the carrier name, the local DNS server, and the outbound traceroute path. The hashed unique device ID provided by the smartphone application development API allows us to distinguish devices while preserving user privacy. Our application also asks users for access permission for their GPS location. So far, this application has already been executed more than 143,700 times on 62,600 distinct devices. *DataSource2* covers about the same time period as *DataSource1*: from September 2009 till October 2010. Given that the application is used globally, we observe a much larger number of carriers, many of which are outside the U.S.

Note that this method of collecting data provides some ground truths for certain data which is unavailable in *DataSource1*, *e.g.*, IP addresses associated with cellular networks instead of Internet end-points via WiFi network can be accurately identified because of the API offered by those mobile OSes.

4.3 Correlating Across Data Sources

One important general technique we adopt in this work, commonly used by many measurement studies, is to intelligently combine multiple data sources to resolve conflicts and improve accuracy of the analysis. This is necessary as each data source alone has certain limitations and is often insufficient to provide conclusive information.

Correlating *DataSource1* and *DataSource2* allows us to tell based on the IP address whether each record in *DataSource1* is from cellular or WiFi networks and recognize the correct carrier names for those cellular records. Under the assumption that a longest matching prefix is entirely assigned to either a cellular network or an Internet wireline network, the overall idea for correlating *DataSource1* and *DataSource2* depicted by Figure 2 is as follows. Both data sources directly provide the IP address information: Each record in *DataSource1* contains the GPS location information reported by the device allocated with the cellular IP address; while *DataSource2* contains the carrier names of those cellular IP addresses. We first map IP addresses in both data sets into their longest matching prefixes obtained from routing table data of *RouteViews*. After mapping cellular IP addresses into prefixes, we have a prefix-to-location table from *DataSource1* and a prefix-to-carrier table from *DataSource2*. Note that the prefix-to-location mapping is not one-to-one mapping because one IP address can be present at multiple locations over time. Combining these two tables results in a prefix-to-carrier-to-location table, which is used to infer the placement of GGSNs after further clustering discussed later.

We believe that cellular network address blocks are distinct from Internet wireline host IP address blocks for ease of management. To share address blocks across distinct network locations requires announcing BGP routing updates to modify the routes for incoming traffic, affecting routing behavior globally. Due to the added overhead, management complexity, and associated routing disruption, we do not expect this to be done in practice and thus assume that a longest matching prefix is either assigned to cellular networks or Internet wireline networks. That is why we map the IP addresses in both data sets to their longest matching prefixes.

Two issues still require additional consideration: (a) building the prefix-to-carrier mapping via *DataSource2*, and (b) evaluating the overlap between *DataSource1* and *DataSource2* to investigate any potential limitation of using *DataSource2* as the prefix-to-carrier ground truth.

4.3.1 Recognizing cellular IP addresses and carriers

We expect *DataSource2* to provide the ground truth for differentiating IP addresses from cellular networks and identifying the corresponding carriers of cellular IP addresses. Each record in *DataSource2* contains the network type, *i.e.*, cellular vs. WiFi, reported by APIs provided by the OS. The carrier name is only available on Android and Windows Mobile due to the API limitation on iPhone OS. After mapping IP addresses to their longest matching BGP prefixes, we can build a table mapping from the BGP prefix to the carrier name for Android and Windows Mobile separately. Although we cannot have a prefix-to-carrier table from iPhone OS, we can produce a WiFi-prefix list tracking all the prefixes reported as WiFi networks, and use this WiFi-prefix list to validate Android's and Windows Mobile's prefix-to-carrier tables. These WiFi prefixes are associated with public IP addresses of the edge networks, likely a DSL or cable modem IP in the case of home users.

We justify the accuracy of Windows Mobile's and Android's prefix-to-carrier tables using the iPhone OS's WiFi-prefix list. We believe iPhone OS's WiFi-prefix list is accurate because there are only limited device types using iPhone OS, *i.e.*, iPhone 4G, iPhone 3G, iPhone 3GS, and iPod Touch, which we tested locally and observed to accurately report the network type. Given a prefix-to-carrier table, we compare it with WiFi-prefix list to detect any potential conflicts, *i.e.*, a case when a prefix in the prefix-to-carrier table appears in the WiFi-prefix list as well. A conflict happens only if one IP address in a BGP prefix is considered as a WiFi address by the *DataSource2* on iPhone OS but listed as a cellular address on Android or Windows Mobile. By comparison, we observe 306 conflicts for Windows Mobile's prefix-to-carrier table, yet no conflict for Android. The reason may be that *DataSource2* on Windows Mobile failed to tell the network type on some platforms since the Windows Mobile OS is customized for each type of phone. Therefore, we use Android's prefix-to-carrier table as the authoritative source for identifying the carrier of each record in *DataSource1* data set.

4.3.2 Overlap between data sources

Set	# BGP prefixes	% in *DataSource1*	% in *DataSource2*
DataSource1 $\cup$ *2*	453	-	-
DataSource1 $\cap$ *2*	259	99.97%	98.96%
$\in$ *DataSource1* $\notin$*2*	181	0.03%	-
$\in$ *DataSource2* $\notin$*1*	13	-	1.04%

Table 3: Overlap between *DataSource1* & *DataSource2*.

Characterizing the overlap between our two data sources helps us estimate the effectiveness of using *DataSource2* to identify the carrier name of *DataSource1*'s cellular prefixes. Moreover, a significant overlap can confirm the representativeness of both *DataSource1* and *DataSource2* on cellular IP addresses as those two data sources are collected independently.

We first compare the overlap between *DataSource1* and *DataSource2*'s records in the U.S. in terms of number of prefixes within the four carriers as shown in Table 3. Although *DataSource1* and

DataSource2 do not overlap much in terms of number of prefixes, *e.g.*, 181 prefixes in *DataSource1* are excluded by *DataSource2*, in terms of number of records the overlap is still significant due to the disappropriate usage of prefixes, *i.e.*, overlapped prefixes contribute to the majority. 99.97% of *DataSource1*'s records are covered by the prefixes shared by both *DataSource1* and *DataSource2*. Therefore, we have high confidence in identifying the majority of cellular addresses based on *DataSource2*. In addition, the big overlap indicates that both data sources are likely to represent the cellular IP behavior of active users well.

5. IDENTIFYING GGSN CLUSTERS

As mentioned in §3, discovering the placement of GGSNs is the key to understanding the cellular infrastructure, explaining the diverse geographic distribution of cellular addresses. This illuminates the important characteristics of cellular network infrastructure that affect performance, manageability, and evolvability. We leverage the information of the geographic coverage of cellular address blocks to infer the placement of each GGSN because those address blocks sharing the similar geographic coverage are likely managed by the same GGSN. In this section, we (1) identify the geographic coverage of the cellular prefixes in *DataSource1*; (2) cluster those prefixes according to the similarity of their geographic coverage; and (3) infer the placement of GGSNs from the different types of clusters. To validate the clustering results we present three validation techniques based on *DataSource2*, DNS request logs from a DNS authoritative server, and `traceroute` probing respectively.

5.1 Clustering Cellular IP Prefixes

On the Internet, an IP address can often provide a good indication of geolocation, albeit perhaps only at a coarse-grained level, as shown by numerous previous work on IP-based geolocation [25, 15, 19, 37]. However, for cellular networks, it is uncertain due to a lack of clear association of IP addresses with physical network locations, especially given the observed highly dynamic nature of IP addresses assigned to a mobile device [5]. In this section, we derive geographic coverage of cellular address blocks in *DataSource1* to study the allocation properties of cellular IP addresses. We have previously described our methodology how to identify cellular addresses and their corresponding carriers in §4: by aggregating IP addresses to prefixes, we can identify the presence of a prefix at different physical locations based on the GPS information in *DataSource1*.

As discussed in §3, we expect that address blocks with similar geographic coverage are likely be subjected to similar address allocation policy. From our data sets, we do observe similarity of geographic coverage present across address blocks. In Figure 3, both /24 address blocks 22 and 5 from Carrier A have more records in the Southeast region. The geographic coverage of these two prefixes is clearly different from the distribution of all Carrier A's addresses in *DataSource1* shown in Figure 3(c), which is influenced by the population density as well as Carrier A's user base. Moreover, we confirm and further investigate the observation in study [5] that a single prefix can be observed at many distinct locations, clearly illustrating that the location property of cellular addresses differs significantly from that of Internet wireline addresses. The large geographic coverage of these /24 address blocks also indicates that users from both Florida and Georgia are served by the same GGSN within this region.

We intend to capture the similarity in geographic coverage through clustering to better understand the underlying network structure. Also, to verify our initial assumption that carriers do not aggregate their internal routes, we repeat the clustering for /24 address blocks instead of for BGP prefixes by aggregating addresses into /24 address blocks. If cellular carriers do aggregate their internal routes, the number of clusters based on /24 address blocks should be larger than that based on BGP prefixes.

The logical flow to systematically study the similarity of geographic coverage is as follows. Firstly, we quantify the geographic coverage. By dividing the entire U.S. continent into N grids, we assign each prefix a N-dimension feature vector, each element corresponding to one grid and the number of records located in this grid from this prefix. As a result, the normalized feature vector of each prefix is the probability distribution function (PDF) of the girds where this prefix appears. Secondly, we cluster prefixes based on their normalized feature vectors using the *bisect k-means* algorithm for each of the four carriers. The choice of N, varying from 15 to 150 does not affect the clustering results, this is because the geographic coverage of each cluster is so large that the clustering results are insensitive to the granularity of the grid size.

The process of clustering prefixes consists of two steps: (i) pre-filtering prefixes with very few records; and (ii) tuning the maximum tolerable average sum of squared error (SSE) of *bisect k-means*. We present the details next.

5.1.1 *Pre-filtering Prefixes with Very Few Records*

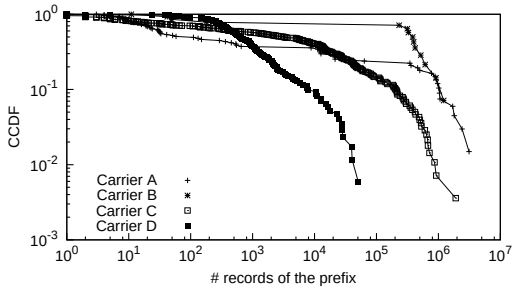


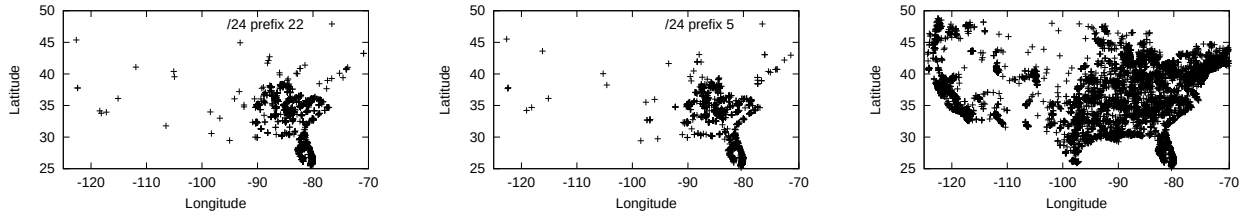
Figure 5: Distribution of # records for prefixes.

Before clustering, we perform pre-filtering to exclude prefixes with very few records so that the number of clusters would not be inflated due to data limitations. Note that aggressive pre-filtering may lead to losing too many records in *DataSource1*.

One intuitive way to filter out those prefixes is to set a threshold on the minimum number of records that a prefix must have. However, the effectiveness of this pre-filtering depends on the distribution of the number of records of prefixes. We plot the complementary cumulative distribution function (CCDF) of the number of records of prefixes in Figure 5. All the four carriers have bi-modal distributions on the number of records of prefixes, implying that we can easily choose the threshold without losing too many records. In our experiments, we choose a threshold for each prefix to be 1% of its carrier's records.

5.1.2 *Tuning the SSE in* bisect k-means *Algorithm*

To compare the similarity across prefixes and further cluster them we use the *bisect k-means* algorithm [35] which automatically determines the number of clusters with only one input parameter, *i.e.*, maximum tolerable SSE. In each cluster, consisting of multiple elements, SSE is the average distance from the element to the centroid of the cluster. A smaller value of SSE generates more clusters. The clustering quality is determined by the geographic coverage similarity of the prefixes within a cluster, which is measured by SSE.



(a). Carrier A's /24 prefix 22 (b). Carrier A's /24 prefix 5 (c). Carrier A's all observed prefixes

Figure 3: Similarity of the geographic coverage for Carrier A's prefixes.

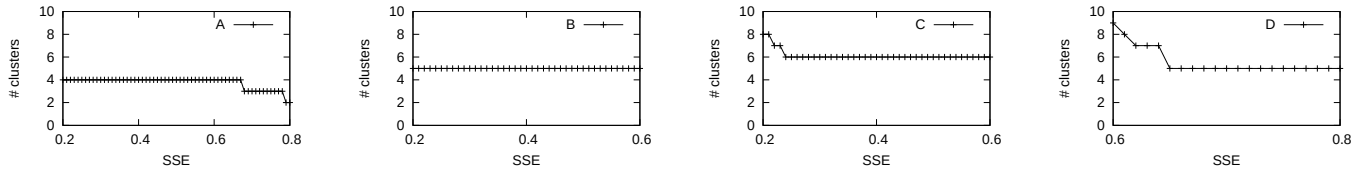


Figure 4: Sensitivity analysis of the SSE in *bi-sect kmeans*.

Figure 4 depicts how SSE, as a measure of the quality of clustering, affects the number of clusters generated for the four carriers. We vary the choice of SSE from 0.01 to 0.99 with increment 0.01. Since there may be multiple stable numbers of clusters, we select the one with the largest range of SSE values. For example, the number of clusters for Carrier A is 4 instead of 3 because it covers [0.2, 0.6] when the number is 4 while it only covers [0.68, 0.78] when the number is 3. From Figure 4, we can also observe that every carriers has an obvious longest SSE range that results in a stable number of clusters, indicating that (i) the geographic coverage across prefixes in the same cluster is very similar; and that (ii) the geographic coverage of the prefixes across clusters is very different.

5.1.3 Clustering Results

We address the problems of pre-filtering and tuning SSE for *bi-sect k-means* clustering in the last two sections. Table 4 shows the parameters we used in pre-filtering and clustering and the clustering results. Aggressive filtering does not happen as every carrier contains at least 99% of the original records after pre-filtering. For Carriers A, B, and C, comparing the clustering at the BGP prefix level vs. the /24 address block level, we do not observe any difference in the number of the clusters generated and the cluster that every address block belongs to. Unlike Carriers A, B, and C, Carrier D does have finer-grained clusters based on its /24 address blocks. We observe that some Carrier D's prefix-level clusters are further divided into smaller clusters at the level of /24 address blocks. These results answer our previous question on the existence of internal route aggregation. Since no internal route aggregation observed for Carriers A, B, and C, BGP prefixes are sufficiently fine-grained to characterize the properties of address blocks. For Carrier D, although the clustering based on /24 address blocks is finer-grained, it does not affect our later analysis. We have applied the clustering on *DataSource1*'s records month by month as well, but we do not see any different numbers of clusters for these 4 carriers.

Figure 6 shows the geographic coverage of each Carrier A's cluster, from the perspective of the U.S. mainland ignoring Alaska and Hawaii, illustrating the diversity across clusters as well as the unexpected large geographic coverage of every single cluster. Note that each cluster consists of prefixes with similar geographic coverage. Each Carrier A's cluster has different geographic spread and center, *i.e.*, Cluster 1 mainly covers the Western, Cluster 2 mainly covers the Southeastern, Cluster 3 mainly covers the Southern and the Mid-Eastern, which are two very disjoint geographic areas, and Cluster 4 mainly covers the Eastern. However, note that the clusters are not disjoint in its geographic coverage, *i.e.*, overlap exists among clusters although those clusters have different geographic centers. For example, comparing Figure 6(b) and 6(d), we can observe that Cluster 2 and Cluster 4 overlap in the Northeast region.

We further quantify the overlap among clusters at grid level. Given a grid, based on all the records located in this grid, we count how many records are from each prefix. Since we know which cluster each prefix belongs to, we can calculate the fraction of records for each grid contributed by different clusters. As a result, for each grid overlapped by multiple clusters, we have a probability distribution function (PDF) on the cluster covering this grid. Based on the PDF, we can calculate the Shannon entropy for each grid. For example, four clusters have 300, 700, 600, and 400 records at grid X respectively, then the PDF for grid X is $[0.3, 0.7, 0.6, 0.4]$ whose Shannon entropy is $-0.3\lg 0.3 - 0.7\lg 0.7 - 0.6\lg 0.6 - 0.4\lg 0.4$. Smaller values of the entropy reflect smaller overlapping degree, *e.g.*, if all the records for a grid are from the same cluster, the grid has an entropy of $-\infty$. Given the number of clusters is N, the theoretical maximum entropy for a grid is $\lg N$.

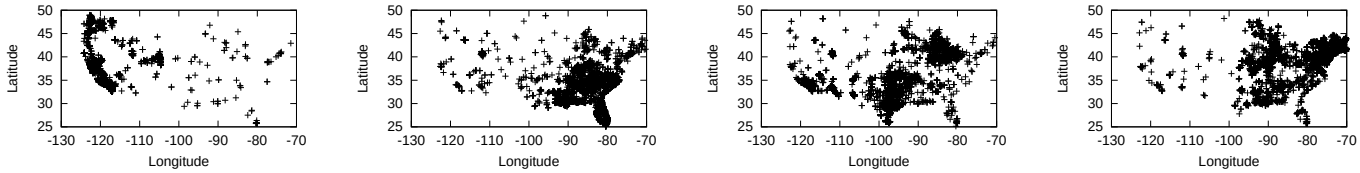
Figure 8 draws the CDF of the entropy of the grid. We can observe that overlap at grid level is quite common for all four carriers, *e.g.*, Carrier A's median entropy value close to 1 means that the records in the corresponding grids are evenly divided by two clusters. We conjecture two reasons for the overlap. The first reason is due to load balancing. Because of user mobility, the regional load variation can be high. Higher overlapping degree is better for maintaining service quality. Moreover, in the extreme case if one cluster has a failure, the overlap can increase the reliability of the cellular infrastructure by shifting the load to adjacent clusters. Another reason is that users commute across the boundary of adjacent clusters. For example, a user in *DataSource1* gets an IP address at a region covered by one cluster, subsequently moves to a nearby region covered by another cluster while still maintaining the data connection. This will result in records showing the overlap between the first and the second cluster in adjacent regions.

Figure 9 shows the clustering results for all four carriers. Al-

Carr.	Thres.		# prefixes		SSE		# clusters		(% of records)[# of prefixes] per cluster	
	BGP	/24	BGP	/24	BGP	/24	BGP	/24	BGP	/24
A	500	300	20	35	0.6	0.5	4	4	(28,19,27,26) [6,5,5,4]	(18,24,25,27) [11,8,8,8]
B	500	300	11	11	0.5	0.5	5	5	(10,14,40,19,17) [1,2,3,2,2]	(10,14,40,17,19) [1,2,3,2,2]
C	500	50	63	245	0.5	0.5	6	6	(28,24,10,7,9,19)[17,11,8,7,6,14]	(50,24,3,3,12,5)[130,59,11,11,23,11]
D[1]	100	100	155	177	0.7	0.2	6	10	(30,10,13,22,9,14) [28,25,28,28,22,24]	(32,6,6,11,6,7,9,4,4,8,4) [27,23,16,16,12,14,11,8,7,7,10]

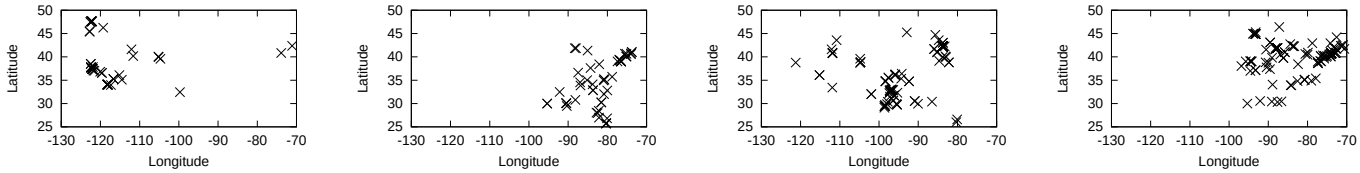
[1] Carrier D's clustering based on /24 address blocks is different with that based on BGP prefixes, which indicates the existence of internal routing

Table 4: Parameters and results for clustering on BGP and /24 address blocks using *bisect k-means*.



(a). Cluster #1 (b). Cluster #2 (c). Cluster #3 (d). Cluster #4

Figure 6: Geographic coverage of each Carrier A's cluster.



(a). Cluster #1 (b). Cluster #2 (c). Cluster #3 (d). Cluster #4

Figure 7: Clustering Carrier A's local DNS servers.

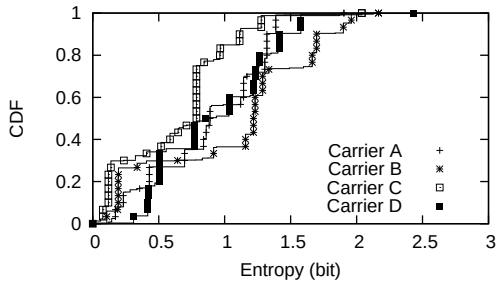


Figure 8: CDF of the entropy of the clusters at the grid level.

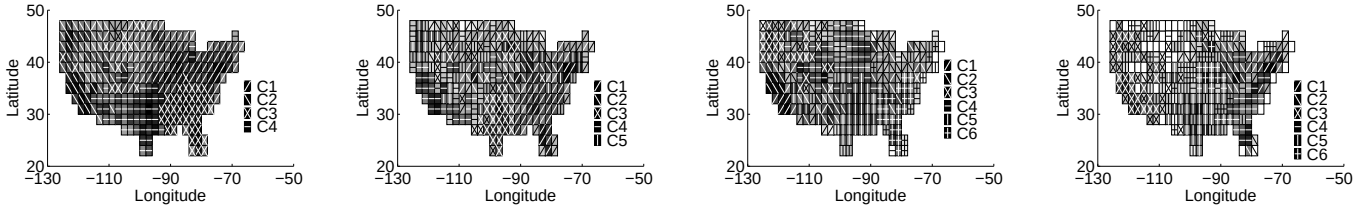
though we have already noticed the overlap among clusters in Figure 6, we are still interested in the dominant geographic coverage of each cluster by assigning every grid to its dominant cluster by majority voting. We make the following observations:

1. All 4 carriers we studied appear to cover the entire U.S. with only a handful of clusters (4–6), each covering a large geographic area, differing significantly from the Internet backbone design.
2. There appears to be some "outlier" cases with sparse presence for each cluster in addition to consistent load balancing patterns. We conjecture that this is caused by limited choice of GGSNs for a small set of devices that use a special set of APNs to which not all GGSNs are available for use.
3. Besides those "outliers", overlap among clusters commonly exists at many locations, *e.g.,* the geographic area around Michigan is clearly covered by three of four Carrier A's clusters. We believe the overlap is due to load balancing and user mobility.
4. Clusters do not always appear to be geographically contiguous. There are clearly cases where traffic from users are routed through clusters far away instead of the closest one, *e.g.*, Carrier A's Cluster 3 covers both the Great Lake area and the Southern region. We believe this is due to SGSNs performing load balancing of traffic across GGSNs in different data centers.
5. The clustering for /24 address blocks is the same as that for BGP prefixes for Carriers A, B, and C confirming that there is no internal route aggregation performed by their cellular IP networks. However, Carrier D has finer-grained clustering for /24 address blocks than that for visible BGP prefixes. Despite this observation, its number of clusters for /24 address blocks is only 10 which is still very limited.

In our analysis, we discover that the infrastructure of cellular networks differs significantly from the infrastructure of wireline networks. The cellular networks of all four carriers exhibit only very few types of geographic coverage. As we expected, the type of geographic coverage reflects the placement of IP gateways. Since the GGSN is the first IP hop, we can conclude the surprisingly restricted IP paths of cellular data network. This network structure implies that routing diversity is limited in cellular networks, and that content delivery service (CDN) cannot deliver content very close to cellular users as each cluster clearly covers large geographic areas.

5.2 Validating Clusters

We validate the clustering result in three independent ways: clustering using *DataSource2*'s records, identifying the placement of local DNS servers in cellular networks, and classifying traceroute paths.



(a). Carrier A (b). Carrier B (c). Carrier C (d). Carrier D

Figure 9: Clusters of all four carriers.

5.2.1 *Validation via DataSource2*

Although the size of *DataSource2* is much smaller than that of *DataSource1*, we can still use *DataSource2* to validate the clustering results obtained from *DataSource1*. We repeat the clustering on the prefixes with more than 100 records from the *DataSource2*. Besides, we repeat the clustering on different types of device, *i.e.,* Android, iPhone, and WM based on *DataSource2*'s records. The clustering results are consistent with those of *DataSource1* in terms of the number of clusters and the cluster that each prefix belongs to. Moreover, all the observations from *DataSource1* listed in §5.1 consistently apply to *DataSource2*.

5.2.2 *Validation via Local DNS Server Based Grouping*

Carrier	# user	# records	# LDNS	# clusters
A	289	384	12	4
B	574	1045	4	1
C	704	884	12	3
D	122	142	15	3

Table 5: Statistics of local DNS experiments.

The configuration of the local DNS infrastructure is essential to ensure good network performance. Besides performance concerns, local DNS information is often used for directing clients to the nearest cache server expected to have the best performance. This is based on the key assumption that clients tend to be close to their configured local DNS servers, which may not always hold [29]. In this work, we perform the first study to examine the placement and configuration of the local DNS servers relative to the cellular users and the implication of the local DNS configuration of cellular users on mobile content delivery. It is particularly interesting to study the correlation between the local DNS server IP and the device's physical location. Since DNS servers are expected to be placed at the same level as IP gateways, *i.e.,* GGSNs, we expect to see similar clusters of cellular local DNS servers based on the geographic coverage.

To collect a diverse set of local DNS server configurations, we resort to our *DataSource2* application by having the client send a specialized DNS request for a unique but nonexistent DNS name which embeds the device identifier and the timestamp (`id_timestamp_example.com`) to a domain (`example.com`) where we have access to the DNS request logs on the authoritative DNS server. The device identifier, `id_timestamp`, is used for correlating the corresponding entry in the *DataSource2*'s log which stores the information such as the GPS information, the IP address, *etc.* The timestamp ensures that the request is globally unique so that it is not cached. This is a known technique used in previous studies for recording the association between clients and their local DNS servers [23]. Since most DNS servers operate in the iterative mode, from the authoritative DNS server, we can observe the formatted incoming DNS requests from local DNS servers.

We summarize our results in Table 5. The four carriers appear to have different policies for configuring local DNS servers. All the local DNS servers of Carrier A across the country fall into one /19 address block. Carrier B altogether only has four distinct DNS IP addresses within two different /24 address blocks, although it has four GGSN clusters. This implies that Carrier B's local DNS servers are unlikely located directly at cellular network gateways, as a single /24 prefix usually constitutes the smallest routing unit. For Carrier C, we observe 12 local DNS server IP addresses within 3 different /24 address blocks. This indicates that, just like Carrier B's clusters, Carrier C's clusters share local DNS servers as well, since Carrier C has more clusters than the /24 address blocks of its local DNS servers. For Carrier D, we observe 15 IP addresses of local DNS servers in 12 /24 address blocks.

For each carrier, we cluster its local DNS servers based on their geographic coverage without any other prior knowledge and show the results in Figure 7. Comparing the clusters based on the local DNS servers with previous clustering based on prefixes in §5.1, we observe that Carrier A's clusters for local DNS servers match very well with the clusters for address blocks (shown in Figure 7). Carriers A's users sharing the same local DNS server IP belong to the same cluster based on cellular prefixes. This serves as another independent validation for previous clustering. Carrier B's users across the U.S. all share the same four local DNS servers, while Carriers C's and D's clusters based on local DNS servers are "one-to-many" mapped to their clusters based on address blocks, indicating that their local DNS servers are shared across multiple clusters as well.

On the current Internet, local DNS-based server selection is widely adopted by commercial CDNs. For Carriers A, C, and D, since their local DNS servers are "one-to-one" or "one-to-many" mapped to GGSNs, server selection based on local DNS servers cannot be finer-grained than the GGSN level. For Carrier B, server selection can be even worse because all Carrier B's local DNS servers are used across the entire U.S.

5.2.3 *Validating via* traceroute *Probing*

Since the clusters created based on cellular prefixes should correspond to the prefixes serving clients within the same network location, we use bi-directional traceroute to further validate this. For the inbound direction, for each prefix of these four carriers in *DataSource1*, we run traceroute on 5 *PlanetLab* nodes at geographically distinct locations within the U.S. to one IP address in this prefix for four days. We make the following observations.

- Stability of traceroute paths at IP level: All traceroute paths obtained from our experiments are found to be very stable without any change at DNS or IP level.
- Stability of traceroute paths at the prefix level: To the same

prefix, the last 5 visible hops in the traceroute path from different *PlanetLab* nodes are consistently the same.

- Similarity of traceroute paths to prefixes in the same cluster: For Carriers A, C and D, prefixes in the same *bisect k-means* cluster share the same traceroute path at DNS or IP level validating their geographic proximity. For Carrier B, each prefix has a distinct traceroute path, making validation more challenging.
- Location correlation between traceroute paths and the cluster's region: For some Carriers A's, B's, and C's clusters, we can infer the GGSN locations from the DNS name of the hops along the path; while for others there is insufficient information to determine router locations. Table 6 shows for the last inferred location along the inbound traceroute path to some clusters with location information inferred from router DNS names. They all agree with the geographic coverage of these clusters.

Cluster	Coverage	DNS key word	Location
A.1	WEST	WA	WA
A.2	SOUTHEAST	GA	GA
A.3	SOUTH	DLSTX	DALLAS, TX
B.1	MIDDLE	CHI	CHICAGO, IL
B.2	SOUTHEAST	FL	FL
B.3	SOUTHEAST	ATLGA	ATLANTA, GA
B.4	WEST	TUSTIN	TUSTIN, CA
B.5	SOUTH	DLSTX	DALLAS, TX
C.1	EAST	CLE	CLEVELAND, OH
C.2	WEST	SCL	SALT LAKE CITY, UT
C.3	NORTHWEST	SEA	SEATTLE, WA
C.4	MIDDLE	AURORA	AURORA, CO
C.5	SOUTH	HOU	HOUSTON, TX
C.6	EAST	NEWARK	NEWARK, NJ

Table 6: Inferred locations for clusters using router DNS names of traceroute paths to the clusters.

Similar to the inbound direction, the outbound traceroute can validate the clustering to some degree. *DataSource2* application runs ICMP traceroute from the device to an Internet server. Assigning the outbound traceroute path to the prefix, we have the following observations:

- For all four carriers, their traceroute paths in the same cluster have the same path pattern, *i.e.*, the sequence of IP addresses or the sequence of address blocks are the same. All clusters are "one-to-one" or "one-to-many" mapped to traceroute path patterns, so each cluster has very different traceroute patterns from the others.
- The prefixes in the same Carrier A's cluster always go through the same set of IP addresses, while for Carriers B, C, and D, their prefixes in the same cluster always go through the same set of /24 address blocks. Therefore we can always tell a prefix's corresponding cluster based on the IP addresses or the /24 address blocks that appear along the traceroute path.

6. IMPLICATIONS ON CONTENT DELIVERY NETWORKS

Based on the previous characterization of cellular data network infrastructure, we highlight the key impact of cellular infrastructure by examining its implication on content delivery networks from the perspectives of content placement and server selection.

6.1 Content Placement

On today's Internet, CDN plays an important role of reducing the latency for accessing web content. The essential idea behind CDN is to serve users from nearby CDN servers that replicate the content from the origin server located potentially far away. By characterizing the cellular infrastructure, we have observed that the current restrictive cellular topology route all traffic through only a handful GGSNs. Therefore, no matter how close to a CDN server the user is, the content still has to go through the GGSN before reaching the destination. The possible reasons for such a restrictive topology design by routing all traffic through GGSNs include simplicity and ease of management, *e.g.*, billing and accounting. Furthermore, it is also easy to enforce policies for security and traffic management. This certainly has negative implication on content delivery.

It is not simple to adapt an existing CDN service, *e.g.*, *Akamai* and *Limelight*, directly to cellular networks due to routing restrictions. One possible alternative is deploying CDN servers within cellular networks to be closer to end users so that the traffic does not have to go through GGSNs to reach the content on the Internet. There has been some startup effort of placing boxes between the RNC and the SGSN to accelerate data delivery and lighten data traffic growth [30], but this design brings additional challenges to management due to the increased number of locations traffic can terminate. Without the support of placing CDN servers inside cellular core networks, placing them close to GGSNs becomes a quick solution for now, and this solution is clearly limited due to the property of the GGSN serving a large geographic region of users.

In *DataSource2*'s application, we measure the ping RTT to 20 Internet servers (landmark servers) located across the U.S. to study the end-to-end latency. The latency to the landmark servers is an approximation on the latency to the content placed at different network locations on the Internet. The 20 servers that we choose are very popular servers geographically distributed across 20 states. To estimate the benefit of placing content close to GGSN, we compare the latency to landmark servers with the latency to the first cellular IP hop, *i.e.*, the first IP hop along the outbound path where GGSN is located.

Each time *DataSource2*'s application runs, it only probes these landmark servers twice to save the resource consumption on devices. In order to eliminate the variability from air interface so that we can isolate the impact from the wireline hops, we follow the splitting method in §5.1 dividing the U.S. continent into N grids. Within each grid, we compare the minimum RTT to the first cellular IP hop against these 20 landmark servers. In Figure 10(a), we show the absolute difference between the latency to the first cellular IP hop and the latency to the landmark servers. Figure 10(b) shows the percentage of the latency saving. Because these 20 landmark servers are widely distributed across the U.S., the minimum latency to landmark servers should be a good estimation of the latency to the current content providers. We can observe that placing content close to the GGSN can reduce the end-to-end latency by 50%. Note that the 50% improvement have already eliminated the variability from air interface. This clearly motivates CDN service providers to push mobile content close to GGSNs.

6.2 Server Selection

Besides the challenge of mobile content placement, server selection is another important issue for CDN service providers. Some existing CDN services, *e.g.*, *Akamai* and *Limelight*, choose the content server based on the incoming DNS requests from the local DNS server assuming the address of the local DNS can accurately represent the location of those end hosts behind the local DNS server. However, this assumption rarely holds for cellular networks.

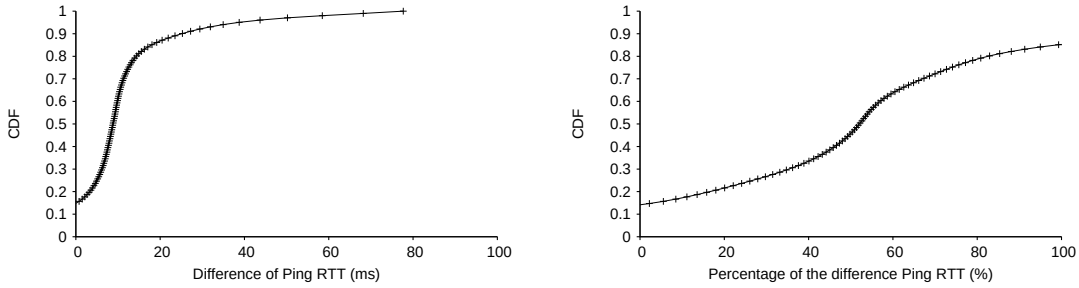


Figure 10: Latency to the first cellular IP hop vs. 20 landmark servers on the Internet.

In §5.2.2, we know that Carriers A, C, and D have different local DNS servers for different GGSN clusters, while Carrier B's clusters share the same set of local DNS servers. Although Carriers A, C, and D have different local DNS server for different GGSN clusters, the IP addresses are very similar. Without the information of the correlation between the local DNS server and the GGSN cluster, it is difficult to choose content servers for different GGSN clusters according to their local DNS server IP address. As Carrier B's GGSN clusters share the same set of local DNS servers, it is impossible to choose content servers for different GGSN clusters based on the DNS request alone.

Interestingly even if content providers can obtain the accurate physical location based on some application-level knowledge, *e.g.*, Google Gears [14], directing the traffic to the content server physically closest to the mobile device can be grossly suboptimal due to the placement of the GGSN and the cellular network routing restrictions. Traffic still needs traverse through the GGSN, despite the close proximity between the mobile device and the content server. To estimate the difference in performance between choosing a server physically closest to the mobile device and one closest to the GGSN node, we do the following analysis. Using the GPS location information reported by *DataSource2*'s application, in all the experiments from Carrier A's Cluster 2, we compare the latency to the landmark server closest to the mobile device with the latency to the landmark server closet to the corresponding GGSN, *i.e.*, one landmark server located at Georgia (according to Table 6 in §5.2.3). Note, similar to §5.1 and §6.1, we split Cluster 2's geographic coverage into grids, aggregate RTTs in the same grid, and compare based on the minimum RTTs as well. Figure 11 shows that the latency to the closet landmark server has high probability to be larger than the latency to the Georgia landmark server and on average by about 10 – 20ms, indicating that choosing the server according to the physical location of the mobile device is suboptimal due to the routing restriction imposed by GGSNs.

Overall, if mobile content providers want to adopt the short-term solution to reduce the end-to-end latency, they have to solve two issues: (i) placing content servers as close as to GGSNs; and (ii) effectively directing traffic to the content server closest to the GGSN that originates the traffic based on information such as the correlation between local DNS servers and GGSNs.

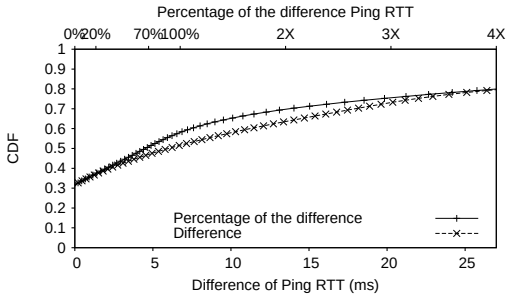


Figure 11: Difference in latency to the closest landmark server from the mobile device vs. to the server closest to the GGSN.

7. CONCLUDING REMARKS

In this paper, we comprehensively characterized the infrastructure of cellular data network of four major wireless carriers within the U.S. including both UMTS and EVDO technology. We unveiled several fundamental differences between cellular data networks and the wireline networks in terms of placement of GGSNs, local DNS server behavior, and routing properties. One of the most surprising findings is that cellular data networks have severe restriction on routing by traversing only a few limited GGSNs to interface with external Internet networks. We observed that all 4 carriers we studied divide the U.S. among only 4–6 GGSNs, each serving a large geographic area. Since the GGSN is the first IP hop, it implies that CDN servers cannot consistently serve content close to end users.

Our study also showed that in the best case local DNS servers for some carriers can be close to GGSNs. Since traffic from and to local DNS servers and cellular users must traverse one of those few GGSNs, using local DNS servers and the knowledge of the mapping to the GGSN to identify the best server to deliver mobile content currently can be sufficient despite the routing restrictions.

Regarding content placement, we investigated and compared two choices: (i) placing content at the boundary between the cellular backbone and the Internet; and (ii) placing content at the GGSN in the cellular backbone. We observed that pushing content close to GGSNs could potentially reduce the end-to-end latency by more than 20%. If pushing content into the proprietary cellular backbone is not permitted, placing content at the boundary still gives considerable benefit.

We believe our findings in characterizing the infrastructure for cellular data networks directly motivate future work in this area. Our observations on the cellular infrastructure guide CDNs to provide better service to mobile users, and our methodology for discovering cellular data network properties will continue to reveal new behavior as cellular networks evolve.

8. REFERENCES

[1] Introduction Mobile Data Track Presentation. `http://www.nanog.org/meetings/nanog47/presentations/Monday/Intro_nanog47_mobiletrack.pdf`.

[2] The future of mobile networking. http://www.nanog.org/meetings/nanog47/presentations/Monday/Future_Mobile_Data_N47_Mon.pdf.

[3] University of Oregon Route Views Archive Project. http://www.routeviews.org.

[4] M. Anand, E. B. Nightingale, and J. Flinn. Self-Tuning Wireless Network Power Management. *Wireless Networks*, 11(4), 2005.

[5] M. Balakrishnan, I. Mohomed, and V. Ramasubramanian. Where's That Phone?: Geolocating IP Addresses on 3G Networks. In *Proceedings of IMC*, 2009.

[6] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In *Proc. ACM SIGCOMM IMC*, 2009.

[7] M. Casado and M. J. Freedman. Peering through the shroud: The effect of edge opacity on IP-based client identification. In *Proc. Symposium on Networked Systems Design and Implementation*, 2007.

[8] R. Chakravorty, S. Banerjee, S. Agarwal, and I. Pratt. MoB: A Mobile Bazaar for Wide-area Wireless Services. In *Proc. ACM MOBICOM*, 2005.

[9] M. C. Chan and R. Ramjee. TCP/IP Performance over 3G Wireless Links with Rate and Delay Variation. In *Proc. of MOBICOM*, 2002.

[10] S. J. Cheng, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt. Constrained Mirror Placement on the Internet. In *Proc. IEEE INFOCOM*, 2001.

[11] J. Chesterfield, R. Chakravorty, J. Crowcroft, P. Rodriguez, and S. Banerjee. Experiences with Multimedia Streaming over 2.5G and 3G Networks. *Journal ACM/MONET*, 2004.

[12] CISCO. Configuring Dynamic Addressing on the GGSN. http://www.cisco.com/en/US/docs/ios/12_4/12_4y/12_4_24ye/cfg/ggsndhcp.html.

[13] M. Ghaderi, A. Sridharan, H. Zang, D. Towsley, and R. Cruz. TCP-Aware Resource Allocation in CDMA Networks. In *Proceedings of ACM MOBICOM*, Los Angeles, CA, USA, September 2006.

[14] Google. Gears. http://gears.google.com.

[15] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based Geolocation of Internet Hosts. *IEEE/ACM Trans. Netw.*, 14(6):1219–1232, 2006.

[16] C. Huang, A. Wang, J. Li, and K. W. Ross. Measuring and Evaluating Large-Scale CDNs. In *Microsoft Research Technical Report MSR-TR-2008-106*, 2008.

[17] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing Application Performance Differences on Smartphones. In *Proc. ACM MOBISYS*, 2010.

[18] E. Inc. LTE-SAE architecture and performance. http://www.ericsson.com/ericsson/corpinfo/publications/review/2007_03/files/5_LTE_SAE.pdf.

[19] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP geolocation using delay and topology measurements. In *IMC 2006: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 71–84, New York, NY, USA, 2006. ACM.

[20] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling Users in a 3G Network Using Hourglass Co-Clustering. In *Proc. ACM MOBICOM*, 2010.

[21] X. Liu, A. Sridharan, S. Machiraju, M. Seshadri, and H. Zang. Experiences in a 3G Network: Interplay between the Wireless Channel and Applications. In *Proceedings of ACM MOBICOM*, 2008.

[22] H. V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani. iPlane: An Information Plane for Distributed Services. In *Proc. Operating Systems Design and Implementation*, 2006.

[23] Z. M. Mao, C. Cranor, F. Douglis, M. Rabinovich, O. Spatscheck, and J. Wang. A Precise and Efficient Evaluation of the Proximity between Web Clients and their Local DNS Servers. In *Proc of USENIX Annual Technical Conference*, 2002.

[24] J. Oberheide, K. Veraraghavan, E. Cooke, J. Flinn, and F. Jahanian. In-Cloud Security Services for Mobile Devices. In *Proc of the First Workshop on Virtualization and Mobile Computing*, 2008.

[25] V. N. Padmanabhan and L. Subramanian. An Investigation of Geographic Mapping Techniques for Internet Hosts. In *ACM Sigcomm*, 2001.

[26] L. Qiu, V. N. Padmanabhan, and G. M. Voelker. On the Placement of Web Server Replicas. In *Proc. IEEE INFOCOM*, 2001.

[27] M. Rahnema. *UMTS Network Planning, Optimization, and Inter-Operation with GSM*. Wiley, 2007.

[28] P. Rodriguez, R. Chakravorty, J. Chesterfield, I. Pratt, and S. Banerjee. MAR: A Commuter Router Infrastructure for the Mobile Internet. In *Proc. ACM MOBISYS*, 2004.

[29] A. Shaikh, R. Tewari, and M. Agrawal. On the Effectiveness of DNS-based Server Selection. In *Proc. IEEE INFOCOM*, Anchorage, AK, April 2001.

[30] S. Solutions. Solutions: Mobile Data Offload. http://www.stoke.com/Solutions/smdo.asp.

[31] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP Topologies with Rocketfuel. In *ACM Sigcomm*, 2002.

[32] N. Spring, D. Wetherall, and T. Anderson. Reverse-Engineering the Internet. In *Proc. First ACM SIGCOMM HotNets Workshop*, 2002.

[33] A. Sridharan, R. Subbaraman, and R. Guerin. Distributed Uplink Scheduling in CDMA Networks. In *Proceedings of IFIP-Networking 2007*, May 2007.

[34] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet Hierarchy from Multiple Vantage Points. In *Proc. IEEE INFOCOM*, 2002.

[35] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[36] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network. In *Proceedings of IMC*, 2009.

[37] B. Wong, I. Stoyanov, and E. G. Sirer. Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts. In *Proc. Symposium on Networked Systems Design and Implementation*, 2007.

[38] M. Zhang, C. Zhang, V. Pai, L. Peterson, and R. Wang. PlanetSeer: Internet Path Failure Monitoring and Characterization in Wide-Area Services. In *Proc. Operating Systems Design and Implementation*, 2004.

[39] Z. Zhuang, T.-Y. Chang, R. Sivakumar, and A. Velayutham. A3: Application-Aware Acceleration for Wireless Data Networks. In *Proc. of ACM MOBICOM*, 2006.