Cell Spotting: Studying the Role of Cellular Networks in the Internet

John P. Rula
Northwestern University/Akamai
john.rula@eecs.northwestern.edu

Fabián E. Bustamante
Northwestern University
fabianb@eecs.northwestern.edu

Moritz Steiner
Akamai
mosteine@akamai.com

ABSTRACT

The impressive growth of the mobile Internet has motivated several industry reports retelling the story in terms of number of devices or subscriptions sold per regions, or the increase in mobile traffic, both WiFi and cellular. Yet, despite the abundance of such reports, we still lack an understanding of the impact of cellular networks around the world.

We present the first comprehensive analysis of global cellular networks. We describe an approach to accurately identify cellular network IP addresses using the Network Information API, a non-standard Javascript API in several mobile browsers, and show its effectiveness in a range cellular network configurations. We combine this approach with the vantage point of one of the world's largest CDNs, with servers located in 1,450 networks and clients distributed across across 245 countries, to characterize cellular access around the globe.

We find that the majority of cellular networks exist as mixed networks (i.e., networks that share both fixed-line and cellular devices), requiring prefix – not ASN – level identification. We discover over 350 thousand /24 and 23 thousand /48 *cellular* IPv4 and IPv6 prefixes respectively. By utilizing addresses level traffic from the same CDN, we calculate the fraction of traffic coming from cellular addresses. Overall we find that cellular traffic comprises 16.2% of the CDN's global traffic, and that cellular traffic ranges widely in importance between countries, from capturing nearly 96% of all traffic in Ghana to just 12.1% in France.

CCS CONCEPTS

• **Networks → Network measurement**; **Mobile networks**;

KEYWORDS

Cellular Networks, Internet Census, Cellular Identification

ACM Reference Format:
John P. Rula, Fabián E. Bustamante, and Moritz Steiner. 2017. Cell Spotting: Studying the Role of Cellular Networks in the Internet. In *Proceedings of IMC '17, London, UK, November 1–3, 2017,* 14 pages.
https://doi.org/10.1145/3131365.3131402

IMC '17, November 1–3, 2017, London, UK

ACM ISBN 978-1-4503-5118-8/17/11.
https://doi.org/10.1145/3131365.3131402

1 INTRODUCTION

The tremendous growth of the mobile Internet, with over 11 billion devices connected by 2020 [8], and its economic implications have motivated several reports retelling the story in terms of number of devices sold, 3G/4G subscriptions per regions or the increase in mobile traffic (both WiFi and cellular) based on survey-based estimations [5, 6, 10, 18]. And yet, despite the abundance of these reports, we lack an understanding of the impact of cellular networks around the world.

There are a number of reasons for this. For one, it is currently challenging to tell whether a particular IP address comes from a cellular or fixed-line user. In much of the world, cellular users reside in networks that combine both cellular and fixed-line customers, complicating any straightforward attempt at identification. Knowing a device type (e.g., smartphone or tablet) has limited value as most mobile devices have multiple interfaces and users tend to offload cellular traffic to WiFi when available. And while instrumented devices or data collected from a network operator's core could provide detailed information on cell network usage, scaling this sort of studies have proven to be difficult [24, 27].

A comprehensive understanding of cellular access has a wide range of applications for different stakeholders in the Internet. For content providers and delivery networks, identifying access technology would help diagnosing and addressing performance issues in the wild. Researchers and operators could better understand how networks are being used around the world and identify potential trends, while policy makers could have a firmer statistical footing for investment decisions.

In this paper we tackle a straightforward yet challenging problem: "Can we estimate the relative importance of cellular networks around the world?" We make two key contributions. First, we describe an approach to accurately identify cellular network addresses using client browser signals and show its effectiveness in a range of mixed networks (i.e., networks which share both fixed line and cellular devices). Using this approach, we leverage the global vantage point of one of the world's largest CDNs to map global cellular IP space and its housing ASes. Our second main contribution is a first-of-its-kind study characterizing cellular network configuration and usage around the world.

A summary of our key findings includes:

- We identify 350 thousand cellular /24 IPv4 subnets, and 23 thousand /48 IPv6 subnets worldwide, and that these comprise 7.3% and 1.2% of active IP address space respectively.
- We identify 668 cellular ASes, and show that a majority (58.6%) of cellular access networks are "mixed networks", housing both cellular and fixed-line broadband customers in the same AS.

Source	Granularity	Global	Comp. Cellular
Industry Reports			
Ericsson [10]	Continent	✓	✓
Cisco [8]	Continent	✓	✓
Sandvine [34, 35]	Continent	✓	X
Akamai SoTI [5]	Country	✓	X
OpenSignal [26]	Country	✓	X
Academic Research			
Flow Analysis [41]	Operator	X	X
Instr. Handsets [12, 13]	Handset	X	X
Our Approach	**IP-Level**	✓	✓

Table 1: Comparison between existing analysis of cellular network usage and behavior.

- We find a high concentration of traffic in a very small fraction of cellular subnets. In a large European operator, 24 out of 514 – thats 4.6% – active cellular /24s account for 99.5% of cellular demand.
- We find that cellular traffic represents 16.2% of all global traffic in December 2016. We show that the fraction of traffic traversing cellular links varies widely across countries and continents. For example, while only 16.6% of U.S. traffic is cellular, cellular composes 63% of all traffic in Indonesia and 95.9% of all traffic in Ghana.
- We find that with a few exceptions – namely the U.S. and India – that IPv6 is not widely deployed across global cellular networks. We found only 1.2% of all active IPv6 /48 subnets are cellular, and are found in only 52 of the 668 (7.7%) cellular ASes.

In the following section we expand on our motivation and describe current approaches for studying and characterizing mobile Internet trends. After describing our datasets (§ 3), we present our method for cellular address identification, and report on its validation and early results in Section 4. We present an approach that builds on these ideas for detecting cellular access networks (§ 5) and apply it to analyze some key features of different cellular networks (§ 6). We discuss some of the observed global trends in Section 7) and close with a summary of our findings and some of its implications. We conclude in Section 8 with some final thoughts and future research directions.

2 BACKGROUND & MOTIVATION

Cellular access technology continues to improve at rapid pace, with existing LTE deployments capable of supporting data rates up to 100 Mbps. The next generation of wireless technology – 5G – is expected to support data rates up to 1Gbps [22]. The improved performance and the proliferation of advanced wireless technologies are driving exponential growth on cellular traffic.

Given its increasing importance, all Internet stakeholders – from users and content providers to content delivery networks, operators, researchers and policy makers – could benefit from a comprehensive understanding of cellular access. Content providers and delivery networks, could better diagnose and address performance issues in the wild. Researchers and operators could better understand how networks are being used around the world and identify potential trends, while policy makers could have a firmer statistical footing for investment decisions.

Distinguishing cellular traffic within more general traffic by mobile devices is a challenging problem. "Mobile" devices describe a property of the device itself, typically a smartphone or tablet, whereas a "mobile" connection describes the type of access connection. We refer to mobile connections strictly as connections traversing cellular access technologies, and focus on the scope and deployment of cellular connectivity.

2.1 Related Work

Information about cellular networks comes from two main sources: academic research and industry reports. Prior academic work on cellular network characterization has typically followed one of two models, either relatively small detailed studies involving instrumented handsets or flow-level analysis from a single mobile operator. Industry reports present high-level analysis and global trends of the current and future Internet, sacrificing specificity for global coverage.

These prior approaches make trade-offs between the coverage of their results and the level of detail of their findings – from "broad and coarse" views found in most industry reports to "narrow and detailed" perspectives collected from instrumented devices. Table 1 presents a summary of these trade-offs within prior work, comparing existing approaches across the granularity of their results, whether they provide a global view, and if they provide a comparative view of cellular and fixed-line traffic.

Instrumented handsets provide the highest level of detail, and today are the only way to obtain all of a device's context, including location and radio conditions. This approach has been used to explore cellular network infrastructure [33, 38, 40], measure performance [16, 23, 36], and understand mobile device behavior [12, 13]. While detailed in their measurements, these approaches' typically limited coverage hampers their ability to observe global trends on global cellular connectivity.

Flow-level analysis from cellular operators allow for more general statements on cellular network behavior since they typically cover orders of magnitude more users over continuous time spans. Several efforts have explored the traffic patterns of cellular networks using this approach, including Shafiq et al. [37] and Zhang et al. [41]. Other work has looked at more specific phenomena, including the dominant share of video in mobile networks [11], or the impact of caching on mobile devices [31]. While providing a in-depth view of a given network, such studies capture only the perspective of a single operators. As we later show (§6) cellular networks vary greatly in their size and configuration.

Often quoted industry reports on the state of the current and future of the Internet, such as Cisco VNI report [8], the Sandvine Global Internet Phenomena Report [34, 35], and Akamai State of the Internet Report [5], now include a mobile component, while some more recent survey focus exclusively on mobile networks (e.g.. Ericsson Mobility Report [10] and OpenSignal's State of Mobile Networks [26]. The majority of these reports rely on proprietary company data and survey data from other sources to explore trends and draw estimations on number of devices, subscriptions or mobile traffic.

Source	Period	/24	/48
BEACON	Dec. 2016 (monthly)	4.7M	1.8M
DEMAND	Dec 24-31 2016 (week snapshot)	6.8M	909K

Table 2: CDN's datasets used for cellular address analysis. The BEACON dataset includes 4.7M /24 blocks and1.8M /48 blocks; the DEMAND dataset includes 6.8M /24 blocks and 909K /48 blocks.

Each of these industry reports attempts to capture the global state of mobile networks, but does so across different axes and granularities of coverage. The Sandvine reports [34, 35], for instance, presents comparisons between mobile and fixed-line access at the application level, but does not compare the magnitude of traffic between these two access types. Others capture only performance characteristics of mobile networks, such as the Akamai [5] and OpenSignal [26] reports. Only the Cisco [8] and Ericsson [10] reports provide information which compares the relative impact of cellular networks on overall Internet traffic, yet, even in these cases only offers a comparison of global aggregates. As we show later in Section 7, the variability of cellular usage varies widely across continents, countries, and even ASes, requiring finer grained analysis for true understanding.

Despite these numerous reports, we are still left without a comprehensive understanding of the impact and magnitude of cellular access networks around the globe. Our approach allows the exploration of large-scale trends across operators and regions, while also providing with information at an IP level of detail. Unlike most industry reports which are irreproducible, and based on proprietary data, our approach is easily replicated by individual network services for analysis across their own clients.

3 DATASET

We leverage the vantage point of one of the largest worldwide CDNs, which receives trillions of requests per day. In particular, we rely on two different information sources from the CDN's monitoring platform: real-user-monitoring beacons (BEACON) and overall platform demand measurements (DEMAND). These sources combine the view from over 200,000 vantage points distributed around the world, and includes data from over 46,000 autonomous systems across 245 countries. Table 2 summarizes key aspects of the datasets.

3.1 BEACON dataset

Our BEACON dataset is derived from logs from Javascript beacons, part of the CDN's Real-User Monitoring system (RUM), and contains information such as the timing and page load information obtained from browser instrumentation (e.g. the Resource Timing API [3]), client information including IP address, and data collected by the Network Information API, which we describe in detail below. We utilize logs collected over a one month period between December 1, 2016 and December 31, 2016.

Beacons are sourced from page loads of CDN customers that have opted-into this RUM system. This limits the visibility of the beacons to clients of participating customers. Additionally, while

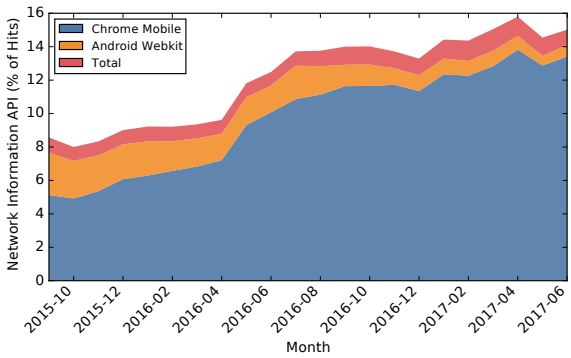


Figure 1: Stacked line graph of the percentage of BEACON hits enabled with the Network Information API. The majority of enabled API hits come from Chrome Mobile and the native Android Webkit browser.

CDNs see a wide variety of HTTP traffic for things such as images, API calls and video streaming, the beacons only capture web page loads. Within this participating customer set, beacons are further sampled from a random subset of all page load requests.

The Network Information API. The Network Information API [2] allows web applications to access information about the underlying network connection in use by the device. While not a W3C standard, the Network Information API is implemented in several popular mobile browsers, most notably Android's native WebKit, Chrome for Android beginning in version 38, [1] and Firefox Mobile.

The API reveals the connection type that the system is using to communicate with the network (e.g., cellular, Bluetooth, Ethernet, WiFi) and supports monitoring network changes. Connectivity is obtained from the browser, which calls the underlying operating system to obtain information on active network interfaces, or to detect changes in network connectivity.

While we have high confidence in the accuracy of Network Information API data (§4), there are issues that arise from classifying access technology type from end-host devices. We discovered two types of issues which lead to inaccurate associations between IP address and connection type. The first and most prevalent is from tethering or mobile hotspot usage. The Network Information API is limited to the device's point of view. Thus, a device that is connected through an intermediate technology, for instance a laptop connected to WiFi through a mobile hotspot – would only report its WiFi network link despite the traffic traversing a cellular network. In another rarer case, there is a possibility that network interfaces could change between when the client's IP address was recorded, and when the Network Information API was polled. In the version of the beacon used in our experiments, client IP addresses were recorded prior to Network Information API invocation, which could lead to this case if a client originally connected to the page over WiFi, and then changed to cellular when the Network Information API was invoked. While it would be possible to monitor

[1] Released on Oct. 8 2014 https://chromereleases.googleblog.com/2014/10/chrome-for-android-update.html

network connectivity changes through the Network Information API to prevent this, the beacon used did not possess this capability. These inaccuracies result in a certain level of noise in the Network Information API's responses, making it unlikely that any heavily trafficked cellular subnet would have a 100% cellular labels. We validate the accuracy of the Network Information API labels in the following section.

Despite its absence in several popular mobile browsers – most notably iOS at the time of our collection – we observe substantial Network Information API traffic in our BEACON dataset. Figure 1 shows the prevalence of Network Information API from our RUM system between September 2015 and June 2017. While at the time of our measurements, the Network Information API was included in 13.2% of beacon requests, that still represents several hundreds of millions of hits. In June 2017, we observe 15% of all BEACON hits to have functional Network Information API data. The figure also shows that the vast majority of Network Information is obtained from Chrome Mobile and Android Webkit browsers, followed by Chrome and Firefox Mobile. Google is heavily driving Network Information API adoption, with 96.7% of enabled requests coming from Google developed browsers in December 2016.

We recognize that our BEACON dataset is biased by the CDN's clientele, as well as the opt-in nature of its RUM system. Regardless of overall biases, we believe our detection methodology is unaffected, since it is relies on the ratio of detected access technology types within individual subnets. We therefore use our BEACON dataset only for network connectivity *identification*, determining whether an IP subnet represents clients connected over cellular or fixed-line access links, and augment our analysis with separate measurements capturing *all* CDN platform demand, across all customers and clients. We describe this additional dataset in detail in the following section.

3.2 DEMAND dataset

We leverage requests logs from the same CDN to generate a comprehensive view of request demand for the entire CDN platform across the global IP space. Using a seven-day period between December 24 and December 31 2016, we develop a platform demand weight for all /24 and /48 subnets which have interacted with the CDN.

Unlike the Javascript beacons which represent a sample of webpage views, these logs accumulate all requests across the CDN's entire platform, covering all types of protocols and devices. To generate this, all daily request statistics are aggregated by /24 subnets for IPv4 and /48 subnets for IPv6. These request statistics are then combined with results from the previous 7 days to smooth out any daily demand variations. Finally, these results are normalized across the platform into unit-less Demand Units (DU). Demand Units are normalized out of 100,000, [2] with each DU representing 0.001% of global request demand (i.e. $1,000DU = 1\%$). These demand records provide a much richer coverage of network demand than the BEACON dataset, and provide context to our results. While a growing fraction of Internet demand is non-web related, much of it continues to operate over HTTP such as video streaming and mobile application traffic.

[2] 100,000 is used to increase precision throughout our analysis

We use demand at this CDN as a proxy for overall traffic demand and acknowledge that this may bias our analysis to areas covered by this particular service. We have, however, no clear way of establishing baselines for mobile traffic usage or assessing sampling bias other than appealing to statistics on the world-wide deployment of the CDN's infrastructure (200,000 servers in 1,450 networks) and the swath of the Internet our requests originate from (46,936 networks in 241 countries).

Compared to the DEMAND dataset, the BEACON dataset captures only 73% of the blocks observed on the entire platform (4.7M out of the 6.4M /24 subnets). The BEACON collection is limited to web page loads, and requires a web browser with Javascript enabled to successfully report data, restrictions that do not apply to the DEMAND dataset collection. When weighting subnets by their respective demand, the BEACON dataset captures 92% of platform requests.

4 CELLULAR SUBNET IDENTIFICATION

In this section we outline our method for cellular subnet identification. The goal is to detect subnets assigned to cellular connection (instead of other type of mobile connections) across the global IP space. We define a cellular connection as one traversing a cellular radio on its path. We then present results from applying this method to our datasets, and report on our validation with ground truth information from three large mobile operators.

4.1 Methodology

Our methodology for classifying subnets as either cellular/non-cellular is straightforward. We use the Network Information API to detect the presence of cellular access technology in a particular IP address block. We compute the ratio of cellular hits for a given subnet, and utilize this ratio to classify each subnet as cellular/non-cellular (i.e., fixed-line). The following paragraphs provide additional details on each of these steps.

To detect the presence of cellular access technology in a particular IP address block we use the `ConnectionType` reported by the Network Information API . `ConnectionType` is defined as an enumeration that includes: Bluetooth, cellular, Ethernet, WiFi and WiMAX.[3]

Using this connectivity information, we label every hit in our BEACON dataset which contains Network Information information as either `cellular` or `non-cellular`, and use this to calculate a *cellular ratio* for every /24 and /48 CIDR sampled. This ratio represents the fraction of a given subnet that comes from `cellular` hits over the total number of Network Interface enabled hits for that subnet. We label a particular subnet as cellular or non-cellular based on this ratio.

Figure 2 plots the distribution of these cellular ratios across global IP space. The figure shows the cumulative distribution of cellular ratios for all active /24 and /48 subnets, as well as the distribution of cellular ratios weighted by these subnets' traffic demand. We find that most addresses fall into two categories: very low cellular (ratio < 0.1) or highly cellular (ratio > 0.9). The figure shows that 91.3% of /24 subnets and 98.7% of /48 subnets in our dataset have a

[3] Other than `WiFi` and `cellular`, all other connection types are rare as the majority of Network Interface enabled browsers operate on mobile devices.

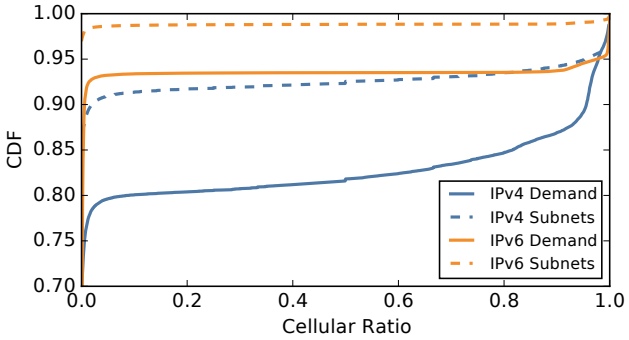


Figure 2: Distribution of calculated cellular ratios for all IPv4 and IPv6 subnets in our BEACON dataset, as well as the traffic demand for these subnets from our DEMAND dataset.

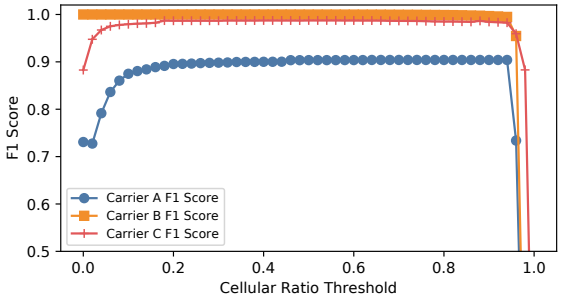


Figure 3: Sensitivity of cellular ratio thresholds for three large mobile operators. We calculated the accuracy, using the F_1 score, of our classifier across different cellular thresholds. The choice of cellular ratio is very resilient, mainly due to the low incidence of false positives of cellular from the Network Information API.

cellular ratio of less than 0.1, and 5.8% of /24 and 1.2% of /48 subnets have a cellular ratio greater than 0.9. The remaining subnets in the range between 0.1 and 0.9 account for 2.9% and 0.1% of /24 and /48 subnets, respectively.

In cases of subnets with intermediate cellular ratios, between 0.1 and 0.9, we have to label them as either cellular or non-cellular assuming access link homogeneity in these aggregates. It is unlikely that ISPs would allocate subnets smaller than /24 and /48 to cellular infrastructure and, indeed, recent studies have found IPv4 /24 subnets to be homogeneous in 90% of cases with respect to last hop routers [19]. We also assume IPv6 /48 subnets to be homogeneous, in light of relative abundance and more recent assignment.

For each subnet in our dataset, we assign a demand value based from the corresponding subnet in our DEMAND dataset. Remember that this dataset accumulates all requests across the CDN's entire platform, normalized into unit-less Demand Units where each units represents 0.001% of global request demand. Looking at demand, we see similar patterns, with the vast majority of subnet traffic residing on either end of the cellular ratio scale. Again, the majority of all traffic demand is contained within subnets with a cellular ratio less than 0.1, making up 80% of IPv4 demand and 98.7% of IPv6 demand. Subnets with cellular ratio greater than 0.9 account for 13.1% of IPv4 demand, and 6.4% of IPv6 demand. There exists, however, substantial demand in the intermediate ratios for IPv4, making up 6.9% of IPv4 demand.

We use a threshold value for cellular ratio to decide on the most appropriate label for a subnet. Clearly, the accuracy of our methodology depends on this chosen threshold. In the following section we describe our process for determining this threshold and validate our choice against the ground-truth from three large mobile operators.

4.2 Parameter Selection & Validation

We derive our threshold values for cellular address identification using ground truth information from 3 large mobile carriers. Our data comes from a diverse set of operators: a large mixed European mobile provider (Carrier A), a large dedicated MNO in the U.S (Carrier B), and a large mixed MNO in the middle east (Carrier C).

For each operator, we obtained a list of IP subnets labeled as belonging to either the cellular or non-cellular (e.g. fixed-line) section of their network. To determine the appropriate threshold for cellular network detection, we compute the accuracy of our method across different threshold values, looking at the precision and recall of our detected cellular subnets compared to ground-truth data.

Precision and recall are common metrics for binary classification. Precision, also called the positive prediction value, is fraction of correctly classified items over the total classified items ($\frac{tp}{tp+fp}$). Recall, also known as sensitivity, is the fraction of correctly labeled items over the true number of items in that class ($\frac{tp}{tp+fn}$). In this context, true positives (TP) represent correctly identified cellular subnets, and true negatives (TN) as correctly identified fixed-line subnets. False positives (FP) denote fixed-line subnets which were identified as cellular, and false negatives (FN) represent cellular subnets which we inaccurately identified as fixed-line.

We compute the accuracy of all threshold values between the range (0,1] by calculating the F1 Score of each threshold. The F1 Score is a combination metric which represents the harmonic mean between precision and recall, and strikes a balance between the accuracy and the comprehensiveness of the classified results. Figure 3 plots this F1 Score threshold sensitivity for all three operators for which we obtained ground-truth data.

The figure shows the stability of classification accuracy across a wide range of threshold values. Across each operator, the accuracy of our detection remains relatively stable for all threshold values between 0.1 and 0.96, implying that our method is robust to different threshold choices. This stability is the result of the high confidence in information that a `cellular` label carries. There are very few cases which lead to cellular false positive results from the Network Information API. Unlike `WiFi` labels which can occur in cellular access links due to tethering and other intermediate connectivity, `cellular` labels are only obtained when the device is connected through a cellular network interface. In this way, even cases where 10% of the reported labels are `cellular` are enough to correctly classify a cellular subnet.

		TP	FP	TN	FN	Precision	Recall	F1
Carrier A^*	CIDR	496	16	89,553	4,626	0.97	0.10	0.09
	Demand	70.96	0.142	1306.36	15.217	0.99	0.82	0.9
Carrier B	CIDR	2,937	0	0	35	1.0	0.99	0.99
	Demand	46.01	0	0	0.016	1.0	0.99	0.99
Carrier C^*	CIDR	383	5	3,049	99	0.98	0.79	0.88
	Demand	10.79	0.17	42.85	0.15	0.98	0.98	0.98

Table 3: Classification accuracy of our approach for a three large mobile operators. Count is the classification accuracy for individual CIDRs, Demand is the classification accuracy weighted by each CIDR's traffic demand. *Mixed operator.

For the remainder of this paper, we set a threshold of 0.5 to denote cellular subnets. We acknowledge that this is a rather conservative threshold given our sensitivity analysis, but we wished to balance the demand curve from Figure 2, which shows low demand below cellular ratios lower than 0.5, and to cover as much cellular demand as possible while minimizing false positives. We believe a simple "majority" label matches these goals.

Validation. We now report on the accuracy of our approach and chosen threshold for cellular network identification. Table 3 reports this accuracy for the three mobile operators, showing the classification category (i.e. *TP*, *FP*, ...) and the *Precision* and *Recall* of our approach. Rows labeled with *CIDR* show the classification accuracy for all active subnets within that operator, and those labeled with *Demand* show the accuracy when weighting CIDRs by their relative traffic demand.

For all the operators, our method produced very high precision, meaning a low false positive rate, for both total CIDRs and the weighted subnet demand. The table highlights the method's high accuracy of classification, showing a precision greater than 0.97 in all instances. This means that for the three operators tested, upwards of 97% of subnets were correctly labeled as cellular. The table also shows that our method misses many no- or low-active cellular subnets, as shown by the large number of false negatives overall. In the case of Carrier A, our approach mislabeled 4,626 cellular subnets as fixed-line subnets. In this way, our approach yields a lower bound on the number of detected cellular subnets, but with a very high confidence in those cellular subnets detected.

4.3 Identifying Cellular Subnets

Applying this methodology to our BEACON dataset, we find a total of 350,687 /24 subnets and 23,230 /48 cellular subnets. South America holds the largest numbers of cellular /24 subnets with 87,589 subnets, closely followed by Asia with 86,618 subnets. North America has only 27,595 /24 subnets despite being one of the largest markets for cellular services.

We find IPv6 deployment within cellular networks to lag significantly behind IPv4. Only 23,230 /48 subnets were detected world wide, and only North America shows substantial deployment of IPv6 addresses. The identified deployment of IPv6 in North American networks corroborates recent findings by Plonka et al. [30] identifying U.S. mobile carriers as some of the largest IPv6 adopters.

Looking at the fraction of addresses that are cellular, we find 7.3% of all active IPv4 /24 prefixes, and 1.2% of IPv6 /48 prefixes to be cellular. We find a wide range in both the numbers and fractions of the IP addresses that are cellular across continents. In Africa and South America, for instance, 53.2% and 22.6% respectively of all /24 subnets detected are cellular. This is in clear contrast to the fraction of cellular subnets found in the remaining continents, which range between 5.7% in Asia to 2.1% in North America. We similarly find a lower relative deployment of IPv6 in cellular networks, again with North America being the exception with nearly 10% of active /48 subnets coming through cellular subnets.

Continent	# /24	# /48	% Active	
			IPv4	IPv6
Africa	79,091	28	53.2%	2.0%
Asia	86,618	4,613	5.7%	0.5%
Europe	65,442	2,117	4.8%	0.3%
North America	27,595	16,166	2.1%	9.9%
Oceania	4,352	35	5.4%	0.07%
South America	87,589	271	22.6%	0.9%
Total	350,687	23,230	7.3%	1.2%

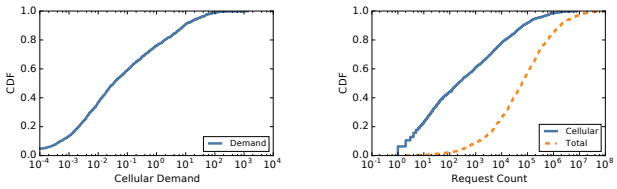
Table 4: Number of detected cellular subnets during December 2016.

Despite the reported benefits of IPv6 in mobile networks, such as improved performance [14], we find IPv6 deployment to be limited across global cellular networks (only 7.7% of operators). In our dataset, we found only 52 of the 668 global networks (7.7%) which support IPv6. Geographically these were found in only 24 countries, with the countries with the greatest numbers of IPv6 networks being Brazil, with 6, and Myanmar, the U.S. and Japan with 5 each. Of those networks, those with the largest numbers of discovered subnet (three out of top four ASes) were in the U.S., and the remaining network in India.

5 CELLULAR AS IDENTIFICATION

In the previous section we applied our methodology to identify cellular network subnets. In the following paragraphs we extend our approach to label ASes. This information is valuable to a variety of services such as content providers and delivery networks, for tasks such as performance debugging, transport customization and the management of performance SLA for their customers, among others.

Using our methodology for subnet identification, a straw-man approach for labeling cellular ASes is to tag any network with 1 or more cellular /24 or /48 subnets as *cellular*. Using such an approach, we find 1,263 (out of 46,936) ASes that fit this category. A cursory



(a) Demand. (b) Beacon hits.

Figure 4: Distribution of demand and beacon responses per ASN.

investigation of the tagged ASes, however, reveals several networks that are obviously not cellular, such as those offering proxy services.

Looking at our initially labeled set, we see that most of the miss-labeled ASes are cloud infrastructure hosting companies or companies offering proxy services for cellular users. Proxy networks for cellular connections are services that reroute traffic from mobile devices, such as performance enhancing proxies for mobile browsers [4]. For example, two of the ASes listed were for Google (AS 15169) and Opera Software (AS 21837), both of which operate performance enhancing proxies for their mobile browsers: Chrome Mobile [4] and Opera Mini [28]. Reverse DNS entries for the proxy cellular addresses corroborate this, having entries such as `google-proxy-*.google.com` for Google's proxy service, and `*.opera-mini.net` for Opera's proxy service. Other common examples include the networks of cloud infrastructure companies, Amazon Web Services or Digital Ocean. We believe these are used to forward traffic from mobile devices for either proxy or VPN services specializing in mobile connectivity [21].

The occurrence of these networks are a product of our data collection approach, which records client IP addresses from the reported analytics beacon data. A connection terminating proxy – which most web accelerating proxies are – will forward the client request through a new HTTP request originating within the proxy's network, yet the connection information contained within the beacon will report the cellular connection actually experienced by the client. A similar problem is experienced by VPN services used by mobile clients, since their external IP addresses are representative of the VPN service. In the following paragraphs we describe several heuristics for refining the preliminary list of cellular-tagged ASes.

5.1 Determining Cellular ASes

To filter out the aforementioned false positives (i.e., from cloud and proxy services) we rely on a set of heuristics. The input to these heuristics is the collection of ASes with one or more detected cellular subnet. The followings paragraphs provides details on such heuristics and their application to the BEACON dataset.

1: Exclude ASes with low cell subnets' demand. From the input set of potential cellular ASes, we find that a large fraction of them have small amounts of overall cellular demand. Figure 4a displays the distribution of demand from each of these 1,263 ASes, showing that 40% of such ASes represent more than 6 orders of magnitude less demand than the largest cellular ASes. We opted for excluding ASes which have a total cellular demand less than 0.1 DU, removing 493 ASes and leaving 770 ASes. We exclude these low demand ASes as their demand may suggest false-positive cases in our detection methodology, without impacting our planned analysis.

Rule	Filtered	Remaining
1. Exclude ASes with a cumulative cellular demand < 0.1 DU	493	770
2. Exclude ASes with < 300 hits	53	717
3. Exclude based on CAIDA AS-classification	49	668
Totally excluded	595	

Table 5: Summary of the application of AS filtering rules. From the 1,263 ASes with at least one cell CIDR, we are left with 668 (≈53%) after applying these rules.

2: Exclude ASes with low beacon responses. We further exclude from this set those with less than 300 beacon responses. The selected ASes fall in the bottom 5th percentile of all ASes with respect to demand. This excludes an additional 53 networks leaving 717 remaining.

3: Exclude non-access ASes. We utilize CAIDA's AS classifications [1] which labels ASes as either Enterprise, Content or Transit/Access. We filter out all ASes which are labeled at Content or had no known class. This filtering removes the remaining non-access networks such as networks hosting performance enhancing proxy services which exhibit large amounts of "cellular" demand. From the previous set of 717 networks, AS-class filtering reduces this to 668 detected cellular ASes.

Table 5 presents a summary of the application of these heuristics to the BEACON dataset. From the 1,263 ASes in the full dataset (out of 49,936 total) with one ore more cellular CIDRs, we exclude 595 (≈47%) in total, after applying all heuristics. In the remainder of this paper, our results and analysis refer to these 668 ASes as the set of active cellular ASes.

	AF	AS	EU	NA	OC	SA
# ASN	114	213	185	93	16	48
Avg./Country	2.6	4.5	4.2	3.9	2.0	4.0

Table 6: Detected cellular ASes by continent.

We summarize the locations, at the continental level, of the remaining 668 ASes in Table 6. We find different numbers of detected cellular ASes per continent, ranging from the 16 in Oceania to the 213 in Asia, although the average per country in each continent show similar patterns with between 2 to 4.5 cellular ASes per country (for this calculation, we only include countries with at least one detected cellular AS). Note that these are averages, and countries with the largest numbers of cellular ASes in their continent greatly exceed those averages (e.g., 13 in India, 17 in Japan, 25 in China, 29 in Russia and 40 in the U.S).

6 THE SHAPE OF CELL NETWORKS

In this section we explore different features of the 668 identified cellular ASes, including their access technology composition, and their demand at the subnet and operator level. We close with a

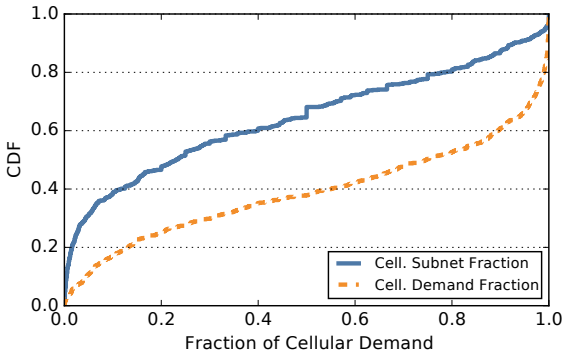


Figure 5: Fraction of cellular demand and cellular subnets for each of the 668 cellular ASes, including mixed networks with both fixed-line and cellular services.

look at DNS resolver usage for cellular clients across these different ASes.

6.1 Mixed Operators

We find that cellular access networks can exist in either dedicated or mixed ASes. We define dedicated ASes as those offering only cellular connectivity to customers, though this can include home broadband delivered over cellular connection. Mixed networks are those offering both cellular and fixed-line services to customers, where fixed-line services include residential broadband technologies like DSL, cable or fiber-to-the-home (FTTH). Here we investigate the degree of mixed networks for global cellular ASes.

Mixed networks arise as many Internet service providers offer both class of services to customers. Through conversations with operators of large mixed networks, we learned that management simplicity and cost savings are some of the main motivations for these networks, especially given the exponential growth in mobile data traffic. While convenient for operators, mixed ASes complicate the work of network services, such as CDNs, trying to optimize performance, or diagnose performance problems of end-users.

We classify cellular networks based on the fraction of their network demand that is cellular. We calculate the *cellular demand (CD)* of an AS as the cumulative demand from all cellular subnets . The *cellular fraction of demand (CFD)* is derived as the ratio of Cellular Demand to the overall demand from all active subnets within that AS.

Figure 5 plots the fraction of each AS's demand that is cellular (CFD). When looking at this distribution, we find no particularly popular configurations of cellular operators, with demand fractions forming a continuous spectrum rather than distinct classes.

To explore this further, we manually investigated the top 50 cellular ASes in terms of cellular demand (CD), labeling each as either *Dedicated* or *Mixed* based on information from the providers' website. In cases where mixed networks use multiple ASes, we still label the AS as cellular if the fraction of demand is greater than 0.95. Using this criteria, we find 32 of the top 50 cellular ASes are dedicated, with the remaining 18 residing in mixed ASes. Looking at the dedicated ASes, we find that 19 of the 32 have demand fractions (CFD) greater than 0.99, and that 29 of the 32 (90%) have fractions greater than 0.95. The lowest fraction of cellular demand in a dedicated operator was 0.9, which we found in an Asian cellular operator. Upon further investigation into why the cellular fraction was so low within a dedicated operator, we found nearly all non-cellular demand contained within two /24 subnets which we believe host HTTP proxies. Each subnet contained substantial platform demand – large numbers of HTTP requests – but almost zero hits in our BEACON dataset – which requires Javascript. We concluded that since no client browsers were active within these subnets, that they were likely terminating proxies. Within the 18 mixed ASes, we find cellular demand comprising anywhere from 4.9% of total demand, up to 81% in certain ASes.

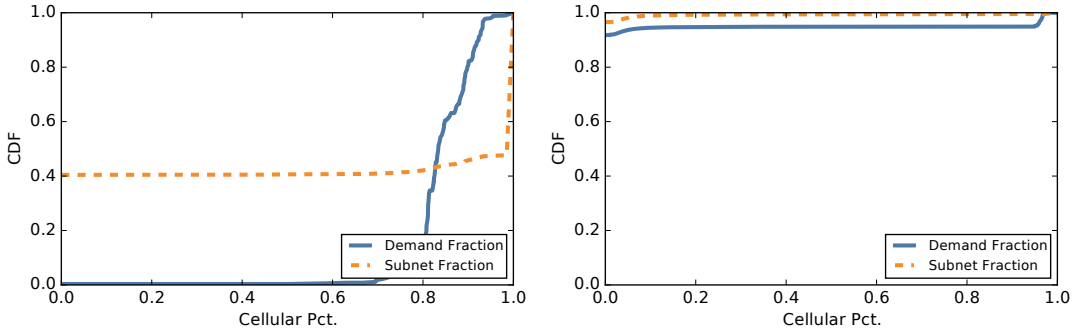
Based on this analysis, we consider any AS with a celular fraction of demand greater than 0.9 to be a dedicated AS, and all those lower than 0.9 to be mixed ASes. Applying this criteria to our dataset, we find that 58.6% of cellular ASes are mixed networks, with 392 mixed to 276 dedicated cellular ASes. We find the locations of mixed operators to be evenly distributed, and roughly half of all detected cellular ASes within all continents. The fraction of mixed operators across continents is relatively equally distributed, with 51% in Africa, 53% in Asia, 56% in Oceania, 61% in Europe, 69% in North America, and 71% in South America. Looking at the demand from each network type, we find that although they outnumber dedicated networks, only 32.7% of cellular demand originates in mixed networks.

Also shown in Figure 5 are the fraction of each AS's subnets that are labeled cellular. We notice the large gap between the distribution of subnet and demand fractions that are cellular. The gap is larger than 0.5 at median, indicating that even in networks where the majority of demand is cellular, a sizable portion of subnets are low-demand and non-cellular. We expand on this disparity between cellular subnetcellular demand in the following section.

Composition of individual mixed networks. We now compare subnet allocation and cellular demand between two large cellular ASes, a dedicated and mixed one. We pick a large U.S. operator as the dedicated cellular AS and a large mixed European operator as the example of a mixed network. The US operator is one of the largest cellular operators in terms of demand. Figure 6 plots, for each cellular AS, a CDF of demand along with a CDF of subnet allocation, across each subnets calculated cellular percentage (§ 4.3)

We see in Figure 6a that even within a dedicated cellular AS, 40% of /24 subnets have a cellular ratio of 0, with virtually no demand. Similar, nearly 50% of addresses with a cellular ratio greater than 0.95 (basically all cellular) also accounts for little to no demand. Nearly all demand in this AS comes from a few /24 subnets which range in cellular ratios between 0.7 and 0.9.

This pattern is in clear contrast with that of mixed operators, which serves fixed-line and cellular customers out of the same AS. Despite being one of the largest cellular providers in its country, less than 2% of its /24 subnets have a cellular ratio greater than 0.2, and capture less than 6% of network demand. In fact, in this operator only 24 /24 subnets account for 99.3% of cellular demand.



(a) Large U.S. dedicated network. **(b) Large European mixed network.**

Figure 6: Breakdown of 2 large cellular ASes, one dedicated and one mixed. The use of active addresses vary widely depending on operator, across overall CIDR space utilization and fraction of cellular demand.

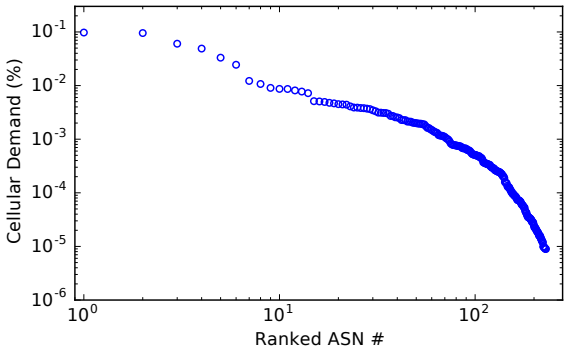


Figure 7: Cellular demand distribution across all identified cellular operators. The top ten global cellular operators hold a disproportionally large fraction of the total, accounting for 38% of global cellular demand.

Rank	Country	Demand (%)	Mixed
1.	US	9.4%	
2.	US	9.2%	
3.	US	5.7%	
4.	IN	4.5%	
5.	US	3.8%	
6.	JP	3.3%	
7.	JP	2.4%	✓
8.	ID	1.5%	
9.	AU	1.2%	✓
10.	JP	1.0%	✓

Table 7: Top ten ASes by demand around the globe.

6.2 Traffic Demand

We now investigate the distribution of cellular traffic demand, first across different operators, and second within individual operator networks.

Traffic demand across operators. We first look at the distribution of demand across our set of global cellular ASes. Figure 7 plots the cellular demand across cellular ASes, with ASes ranked by demand. We represent demand as the normalized fraction of overall global cellular demand originating from each AS. We observe a disproportionally high share of demand contained within the top 10 ASes, and particularly among the top 5 ASes. In fact, these top five ASes alone account for 35.9% of the global cellular demand.

In Table 7 we take a closer look at these top ten ASes, their country of origin and their cellular demand. Even within these top ten operators, traffic is largely skewed towards the very top operators, with the largest mobile AS containing 8.8x the demand from the 10th ranked operator. We first notice that the list is dominated by the two largest cellular ASes with respect to demand, approximately equivalent in their total demand, and each with 62.2% and 61.7% greater demand than the third ranked operator.

Additionally, we can see that these large ASes are located largely in either the U.S. or Japan, which account for 7 out of the top 10 cellular ASes. The U.S. alone constitutes all top three cellular ASes, as well as 4 out of the top 5 ASes. Last we see that while the all top 6 ASes are dedicated cellular, 3 out of the top 10 are mixed operators, meaning they exist in networks composed of both cellular and fixed-line access technologies.

Subnet Traffic Demand. Changing our focus to subnets, we find that cellular traffic is dominated by a small number of /24 subnets. These *heavy-hitter* cellular subnets are much more concentrated in their demand than are seen in fixed subnets. Figure 8 illustrates this for a large mixed European ISP.

In the figure, the majority of cellular demand is distributed across only 25 individual /24 subnets, which capture 99.3% of all cellular demand. After those top 25 subnets, demand in the next largest cellular subnet steeply drops by nearly two orders of magnitude. In contrast, the fixed-line demand is more gradually distributed across its addresses. The drop off in fixed-line subnet traffic occurs after 3 orders of magnitude more addresses than for cellular. In this particular network, cellular demand accounted for only 4.9% of the total, and yet all of the 25 top cellular subnets originated more demand than the largest fixed-line subnet. This can be at least partially explained in light of the widespread use of carrier-grade

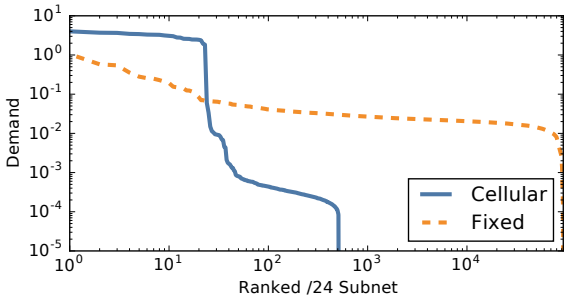


Figure 8: Distribution of subnet demand for cellular and fixed subnets within a large European mixed network. Cellular demand is concentrated within a small handful of /24 subnets, with nearly all cellular demand contained within 25 /24 prefixes. Notice log-scale on the y-axis.

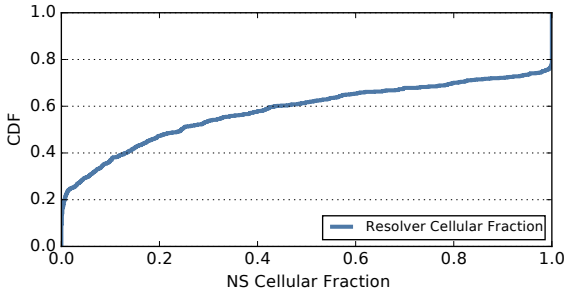


Figure 9: Demand fraction of cellular traffic on cellular resolvers. Nearly 60% of resolvers in mixed cellular networks are shared between cellular and fixed-line customers.

NATs within cellular networks [32, 39]. In the larger context, this means that cellular addresses are some of the most concentrated network subnets on the Internet with regards to demand and that, in many instances, they can be represented by a relatively few number of IP addresses.

6.3 DNS and Cellular Networks

Last we present the results of our analysis of DNS resolvers usage in cellular networks. Our vantage point at a large CDN allows a comprehensive look at the DNS traffic from networks worldwide. In the following paragraphs we look first at the assignment of DNS resolvers in mixed cellular networks and then analyze the use of public DNS services across all cellular networks.

To analyze resolver usage across cellular clients, we first generated client-to-resolver affinities, produced by a similar method to those used by Chen et al. [7]. The end result is a weighted association between client subnet and resolver IP addresses. For our analysis, we combine these client to resolver associations with the previous two datasets, to calculate the amount of demand from each subnet assigned to each resolver. After aggregating these data sources by resolver, we are left with cellular and fixed-line demand originating from each resolver.

Mixed Network Resolvers. We first calculate the fraction of cellular demand across all DNS resolvers in the 392 previously determined mixed cellular ASes. The sharing of resolvers between mobile and fixed-line customers has clear implications for content providers and delivery networks, since DNS-based redirection remains the dominant method for content request routing [33]. Figure 9 plots the CDF of this fraction of cellular traffic across all resolvers in the mixed cellular networks identified in the previous section. A fraction of 0 indicates a resolver that sees only fixed-line requests while a fraction of 1 indicates a resolver that sees *only* cellular traffic requests.

The figure clearly shows that a majority of resolvers - close to 60% - are shared between cellular and fixed-line customers, with the median resolver serving approximately 25% cellular and 75% fixed-line demand. The reminding resolvers appeared to be split evenly (≈ 20% on each side) between those dedicated to cellular and non-cellular requests.

Sharing resolvers between cellular and fixed-line customers may not indicate any issues, since both fixed-line and cellular clients could reside in the same geographic areas, and peer at the same locations. However, we discovered several operators where cellular clients were assigned to distant shared resolvers, yet resolvers were proximal to their fixed-line customers. For example, in a large mixed cellular operator in Brazil, cellular clients in Fortaleza in northern Brazil are assigned to DNS resolvers in Sao Paulo, 1470 miles away. On the other hand, the fixed-line customers assigned to those resolvers were nearly all in Sao Paolo, and represented 80% of that resolvers' end-user demand.

Public DNS Usage. We next look at public DNS usage across cellular networks. Public DNS services have grown increasingly popular in recent years due to claims of greater reliability and their potential for censorship avoidance. On the other hand, previous work has also shown that their use may result in suboptimal redirections to replicas located far away from clients [9, 29]. Despite previous reports to the contrary [33], we find that outside of the U.S. there is significant adoption of public DNS services in cellular operators. To calculate the rate of public DNS usage in cellular networks we use the same methodology as above and compute the fraction of demand resolved through common public DNS services: GoogleDNS [15], OpenDNS [25], and Level3 [20].

Figure 10 shows the fraction of requests coming through three popular public DNS services. While U.S. operators adhere to conventional wisdom on the use of public DNS in mobile operators, with less than 2% of requests being sent through public resolvers, we found a sizable number of global MNOs reliant on public DNS infrastructure despite their potential impact on network performance and QoE. Note that unlike in broadband networks, where users may change their DNS configuration independently of their operator, to use of public DSN service in cell networks implies operator adoption. In one large operator in India, for instance, we see public resolver being use in nearly 40% of cases. Both Honk Kong operators use public resolvers for over 55% of requests, and in the extreme example, we see 97% of request demand coming through public DNS resolver for a mobile operator in Algeria. The latter is most likely due to that operator utilizing a DNS forwarder towards these public options.

Continent	Cellular Fraction (%)	Global Cellular (%)	Subscribers (M) [17]	Demand/ 1000 Subscribers
Oceania	23.4%	3.0%	43.3	**0.0113**
Africa	25.5%	2.9%	954	0.0005
South America	12.5%	4.1%	499	0.0013
Europe	11.8%	15.9%	968	0.0026
North America	16.6%	35%	594	0.0095
Asia*	**26.0%**	**38.9%**	**2,766**	0.0022
Overall	16.2%	100%	5,825	0.0053

Table 8: Cellular demand statistics by continent (* excluding China).

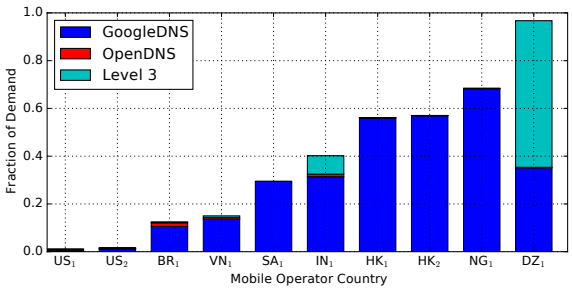


Figure 10: Public DNS usage in selected cellular networks around the globe, labeled by network country. Outside of U.S. operators, we see large fractions of cellular users reliant on public DNS services.

One possible explanation for the prevalence of public DNS usage in operators from countries such as India, China and Brazil may be the reported larger fraction of non-handset devices connected to cellular networks – connected either directly or tethered through mobile hotspot. Investigating the variety of device types across different networks is left as future work.

6.4 Summary of Key Findings

We now summarize the key findings from this section. We presented the first survey on the composition and traffic dynamics of global cellular networks. We highlighted the diversity of the network composition – mixed or dedicated ASes – and supporting infrastructure, such as how network infrastructure such as DNS resolvers are assigned and shared.

Finding 1: A majority of cellular networks (58.6%) are mixed, hosting both cellular and fixed-line broadband clients. Given the prevalence of mixed ASes, efforts on network characterization efforts should take access technology composition into account.

Finding 2: Cellular demand is centralized in a few, large networks. We find that the top 10 cellular ASes account for 38% of global demand. Each of the top 10 operators reside in countries that cover large geographic areas, and have well developed cellular infrastructure. In the future, overall demand may become more equitable as more operators upgrade from 3G to LTE infrastructure.

Finding 3: Cellular traffic is concentrated in a small fraction of IP addresses. Due to the presence of NATs for cellular clients in many networks, a handful of /24 subnets generate much of the cellular demand, while the majority of active cellular addresses carry very little demand. Attempts at measuring cellular networks from IP addresses must be aware of these concentrations in traffic distributions. Such concentrations may also make possible to capture *representative* samples of measurements for these networks with a relatively small number of target addresses.

Finding 4: In mixed cellular networks, nearly 60% of DNS resolvers are shared between cellular and fixed-line clients. This implies that DNS resolvers alone are insufficient for identifying client type. The use of shared resolvers may also challenge client localization for common request routing systems. For instance, we found in a large Brazilian MNO that while cellular and fixed-line clients shared resolvers, the resolvers were only geographically proximal for fixed-line clients. Cellular clients were located over 1470 miles away.

Finding 5: We find significant public DNS usage by cellular clients outside the U.S. This breaks from common assumptions that cellular clients only use operator provided DNS and further complicates attempts at using DNS for end user mapping in CDNs.

7 MACROSCOPIC VIEW OF CELLULAR NETWORKS

In the following section, we take a macroscopic view of cellular networks to identify trends that may help inform the different Internet stakeholders – mobile operators, content providers and policy makers.

7.1 Global Cellular Demand Distribution

We first take a look at the overall distribution of global cellular demand at the continent level. We look at cellular network trends across the following metrics: (*i*) the percentage of a continent's demand that is cellular (*ii*) the percentage of global cellular demand originating from that continent (*iii*) the number of mobile subscribers, [4] per continent [17] and (*iv*) the cellular demand per subscriber. Table 8 summarizes these metrics by continent.

Column 1 presents the percentage of each continent's demand that originates from cellular access links. Overall we find that cellular networks account for 16.2% of *all* demand worldwide. At first take, this appears to be quite different (2-3x) that what has been reported previously by industry. For instance, the 2016 [5] report from Ericsson [10], based on 2015 data, finds mobile traffic to account

[4]Subscriptions refer to all mobile subscriptions including voice, not only mobile data
[5]The latest report to include both mobile and fixed-line data traffic.

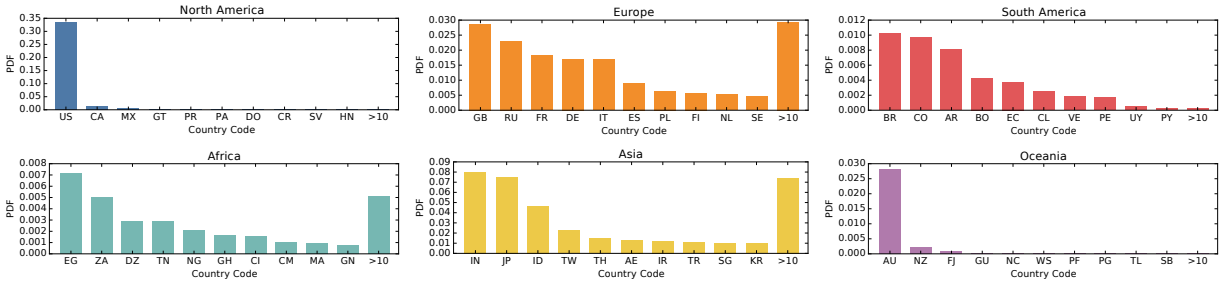


Figure 11: Normalized fraction of global cellular demand by country across continents.

for 8.11% of the global total while the 2017 Cisco VNI states that mobile traffic accounts for 8% of global traffic [8].

The difference can be at least partially explained by our use of *request demand* to calculate demand. While this metric is useful for comparison, e.g. between continents or countries, it is difficult to infer from it overall traffic demand as defined in those reports since, in many cases, objects served for the same application-level request over a cellular connection are much smaller than on a wired connection.

Looking at individual continents, the fraction of cellular demand varies between 11.3% in Europe to 25.5% and 26% in Africa and Asia, respectively. The high fractions of cellular demand in Africa and Asia is partially explained by the limited deployment of fixed-line telecommunications infrastructure. The European's fraction is, however, somewhat surprising given the mature mobile telecommunications industry. It is important to note that our calculations exclude demand data from China. We did not feel confident in the demand values we obtained for Chinese end users, and therefore excluded it from our calculations. Our data therefore underestimates the overall demand fraction from Asia, since it lacks data for China and its 1.3 billion mobile subscribers [17].

We next look at the percentage of cellular demand contained within each continent. We find Asia generates the largest amount of cellular demand, with 38.9% of the global cellular total. North America is the next largest with 35%, followed by Europe with 15.9%. South America (4.1%), Oceania (3%) and Africa (2.9%) compose the remaining 10%. These trends are in line with industry reports [10] showing Asia, followed by North America as the main contributors.

When looking at demand per subscriber (Col. 5), we find that Oceania has the greatest demand per subscriber, followed closely by North America. This may be due in part to the type of access technology, devices and particular subscriptions with large penetration of high-end devices and well-built WCDMA and LTE infrastructure [10]. 4G users have been shown to generate 10x the traffic than 3G networks [8]. Conversely, Africa has the lowest per subscriber demand, which one would expect given its infrastructure is dominated by second generation wireless technology (GSM and EDGE), and has the lowest penetration of 4G wireless technology [10].

7.2 Country-level Statistics

We look next at the distribution of cellular demand across individual countries within each continent. Figure 11 plots the top ten countries within each continent, displaying the fraction of global

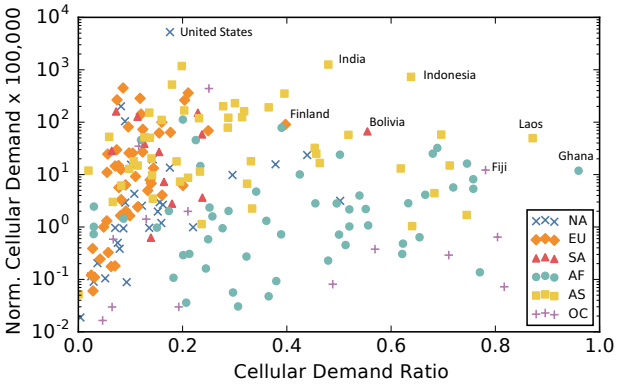


Figure 12: Countries shown in relation to their overall cellular demand (CD) and fraction of cellular traffic (CFD). Notice the log scale on the y-axis.

cellular demand within them. The figure highlights the large impact of top countries, in terms of demand, on the overall distribution of cellular demand. We observe a clear heavy tail distribution, with the U.S. accounting for over 30% of global cellular demand. The top 5 countries, all in Asia with the exception of the U.S., contribute 55.7% of global cellular demand, and the top 20 countries make up a notable 80% of the global total.

We investigate these trends further by looking, for each country, at the relationship between cellular demand and the fraction of that country's total demand labeled as cellular. Figure 12 plots this relationship with the country cellular demand (CD) on the y axis and the fraction of total demand that is cellular (CDF) on the x axis.

The figure shows the wide range of degrees in which cellular connectivity is relied on in different countries. We find the majority of countries within Europe, South and North America all clustered together on the far left, composed of cellular fractions of demand lower than 0.2. The right 80% of the figure is populated mainly by countries in Africa and Asia, and represent cellular dominant network connectivity.

The frontier on the upper right side of the figure shows countries with either very high levels of cellular demand (e.g., US), a very high fraction of overall cellular traffic (e.g., Ghana), or both (e.g., Indonesia). For instance, although the U.S. has by far the largest overall cellular demand, it only accounts for 16.6% of overall country

traffic. In contrast, cellular demand in Laos and Ghana represents 87.1% and 95.9% of their overall country demand respectively. On both extremes, Indonesia (ID), the 4th largest country for cellular demand, uses cellular connectivity for 63% of its country's traffic demand! We believe these countries along the frontier represent ideal targets for further study to understand the traffic dynamics and user behavior in areas which are *already* mobile dominated.

7.3 Summary of Key Findings

We summarize the key findings from our macroscopic view of cellular networks. This perspective illustrates the large dominance in terms of traffic of a few markets, such as the U.S, and the various roles played by cellular connectivity around the globe – from a supplementary service in much of Europe to the primary means of connectivity in Asia and Africa.

Finding 1: Cellular traffic makes up 16.2% of all global traffic demand in our dataset. In Asia and Africa, cellular traffic accounts for 25.5% and 26% of continent demand, respectively.

Finding 2: In terms of cellular traffic demand, the top countries dominate the overall distribution, to an even greater degree than AS level demand. The top 5 countries account for 55.7% of global cellular traffic demand, and the top 20 comprise 80%.

Finding 3: In several countries (e.g., Laos, Indonesia), cellular access is the dominant form of Internet connectivity and thus, increasingly, part of these countries' critical infrastructure.

8 CONCLUSION

This paper presents the first global analysis of cellular networks. We described an approach to accurately identify cellular network addresses using client browser signals and showed its effectiveness in a range of mixed networks (i.e., networks that share both fixed-line and cellular devices). Using this approach, we leveraged the global vantage point of one of the world's largest CDNs to map the global cellular IP space and their hosting ASes, and analyzed their traffic demand.

There are several directions we would like to explore in future work. Though this paper presented a snapshot of cellular address characteristics, we are exploring how cellular addresses evolve over time, both in their assignment to cellular end-users, and how demand shifts across cellular address space. In addition, we would like to use our new map of cellular addresses, and our global CDN vantage point to characterize user behavior across a wide range of network services.

ACKNOWLEDGMENTS

The authors would like to thank our shepherd Aaron Schulman and the anonymous IMC reviewers for their valuable comments and helpful suggestions. We also want to extend our thanks to Pablo Alvarez and K.C. Ng, for their insight and help throughout this process. This work was partially supported by NSF CNS-1218287.

REFERENCES

[1] The caida ucsd as classification dataset, 08-01-2015. http://caida.org/data/as_classification.xml.
[2] Network information api. https://wicg.github.io/netinfo/.
[3] Resource Timing API Level 3. https://w3c.github.io/resource-timing.
[4] V. Agababov, M. Buettner, V. Chudnovsky, M. Cogan, B. Greenstein, S. McDaniel, M. Piatek, C. Scott, M. Welsh, and B. Yin. Flywheel: Google's data compression proxy for the mobile web. In *Proc. USENIX NSDI*, 2015.
[5] Akamai. State of the internet. https://www.akamai.com/uk/en/our-thinking/state-of-the-internet-report/.
[6] M. M. K. P. C. Byers). Internet trends 2016. http://www.kpcb.com/blog/2016-internet-trends-report, June 2016.
[7] F. Chen, R. K. Sitaraman, and M. Torres. End-user mapping: Next generation request routing for content delivery. In *Proc. ACM SIGCOMM*, 2015.
[8] CISCO. CISCO global mobile data traffic forecast update, 2016-2021 white paper. Technical report, CISCO Systems Inc., 2016.
[9] C. Contavalli, W. van der Gaast, D. Lawrence, and W. Kumari. Client subnet in dns queries, 2016.
[10] Ericsson. Ericsson Mobility Report: June 2016. Technical report, Ericsson, 2016.
[11] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck. Over the top video: the gorilla in cellular networks. In *Proc. IMC*, 2011.
[12] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A first look at traffic on smartphones. In *Proc. IMC*, 2010.
[13] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *Proc. of MobiSys*, 2010.
[14] U. Goel, M. Steiner, M. P. Wittie, M. Flack, and S. Ludin. A case for faster mobile web in cellular ipv6 networks. In *Proc. of MobiCom*, 2016.
[15] Google. Frequently asked questions - Public DNS – Google Developers. https://developers.google.com/speed/public-dns/faq.
[16] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proc. of MobiSys*, 2012.
[17] ITU. Statistics - mobile-cellular subscriptions. http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.
[18] M. Kende. Internet Societ Global Internet Report 2015: Mobile Evolution and Development of the Internet. http://www.internetsociety.org/globalinternetreport/2015/assets/download/IS_web.pdf, 2015.
[19] Y. Lee and N. Spring. Identifying and aggregating homogeneous ipv4/24 blocks with hobbit. In *Proc. IMC*, 2016.
[20] Level 3. Level 3 Communications. http://www.level3.com/.
[21] A. Molavi Kakhki, A. Razaghpanah, A. Li, H. Koo, R. Golani, D. Choffnes, P. Gill, and A. Mislove. Identifying traffic differentiation in mobile networks. In *Proc. IMC*, 2015.
[22] NGMN Alliance. NGMN 5G White Paper. Technical report, Next Generation Mobile Network Alliance, 2016.
[23] A. Nikravesh, D. R. Choffnes, E. Katz-Bassett, Z. M. Mao, and M. Welsh. Mobile network performance from user devices: A longitudinal, multidimensional analysis. In *Proc. PAM*, 2014.
[24] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao. Mobilyzer: An open platform for controllable mobile network measurements. In *Proc. of MobiSys*, 2015.
[25] OpenDNS. Cloud Delivered Enterprise Security with OpenDNS. https://www.opendns.com.
[26] OpenSignal. Global state of mobile networks. https://opensignal.com/reports/2016/08/global-state-of-the-mobile-network.
[27] OpenSignal. Opensignal: 3g and 4g lte cell coverage map. https://opensignal.com.
[28] Opera. Opera mini - mobile browser with an ad blocker. http://www.opera.com/mobile/mini.
[29] J. S. Otto, M. A. Sánchez, J. P. Rula, and F. E. Bustamante. Content delivery and the natural evolution of DNS: Remote DNS trends, performance issues and alternative solutions. In *Proc. IMC*, 2012.
[30] D. Plonka and A. Berger. Temporal and spatial classification of active ipv6 addresses. In *Proc. IMC*, 2015.
[31] F. Qian, K. S. Quah, J. Huang, J. Erman, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck. Web caching on smartphones: ideal vs. reality. In *Proc. of MobiSys*, 2012.
[32] P. Richter, F. Wohlfart, N. Vallina-Rodriguez, M. Allman, R. Bush, A. Feldmann, C. Kreibich, N. Weaver, and V. Paxson. A multi-perspective analysis of carrier-grade nat deployment. In *Proc. IMC*, 2016.
[33] J. P. Rula and F. E. Bustamante. Behind the curtain: Cellular dns and content replica selection. In *Proc. IMC*, 2014.
[34] Sandvine. 2016 global internet phenomena: Africa, asia-pacific and middle east. Technical report, Sandvine Incorporated ULC, 2016.
[35] Sandvine. 2016 global internet phenomena: Latin america & north america. Technical report, Sandvine Incorporated ULC, 2016.
[36] S. Sen, J. Yoon, J. Hare, J. Ormont, and S. Banerjee. Can they hear me now?: A case for a client-assisted approach to monitoring wide-area wireless networks. In *Proc. IMC*, 2011.
[37] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *Proc. ACM SIGMETRICS*, 2011.
[38] N. Vallina-Rodriguez, S. Sundaresan, C. Kreibich, N. Weaver, and V. Paxson. Beyond the radio: Illuminating the higher layers of mobile networks. In *Proc. of MobiSys*, 2015.
[39] Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang. An untold story of middleboxes in cellular networks. In *Proc. ACM SIGCOMM*, 2011.

[40] Q. Xu, J. Huang, Z. Wang, F. Qian, and A. G. Z. M. Mao. Cellular data network infrastructure characterization and implication on mobile content placement. In *Proc. ACM SIGMETRICS*, 2011.

[41] Y. Zhang and A. Årvidsson. Understanding the characteristics of cellular data traffic. *ACM SIGCOMM Computer Communication Review*, 42(4):461–466, 2012.