Geolocating IP Addresses in Cellular Data Networks

Sipat Triukose[◊], Sebastien Ardon[◊], Anirban Mahanti[◊], Aaditeshwar Seth[‡]

[◊] NICTA, Locked Bag 9013, Alexandria, NSW, Australia
{sipat.triukose,sebastien.ardon,anirban.mahanti}@nicta.com.au
[‡] IIT Delhi, New Delhi, India
aseth@cse.iitd.ernet.in

Abstract. Smartphones connected to cellular networks are increasingly being used to access Internet-based services. Using data collected from smartphones running a popular location-based application, we examine IP address allocation in cellular data networks, with emphasis on understanding the applicability of IP-based geolocation techniques. Our dataset has GPS-based location data for approximately 29,000 cellular network assigned IP addresses in 50 different countries. Using this dataset, we provide insights into the global deployment of cellular networks. For instance, we find that Network Address Translation (NAT) is commonplace in cellular networks. We also find several instances of service differentiation with operators assigning public IP addresses to some devices and private IP addresses to other devices. We also evaluate the error of geolocation databases when determining the position of the smartphones, and find that the error is 100km or more for approximately 70% of our measurements. Further, there is potential for errors at the scale of inter-country and inter-continent distances. We believe this dataset may be of value to the research community, and provide a subset of the dataset to the community.

1 Introduction

Estimating the geographical location of Internet hosts has many applications including targeted marketing, user profiling, fraud detection, regulatory compliance, digital rights management, and server or content distribution network performance tuning. For instance, to comply with region-specific licensing arrangements, many streaming media services restrict content access based on the user's geographic location. One popular approach to geolocation is the use of database services such as Maxmind [2] and IPinfoDB [1] that maintain an exhaustive table of IP prefix to location matches. However, dynamic assignment of IP addresses, increased fragmentation of IP address blocks, and extensive use of middleboxes make IP-based geolocation extremely challenging.

In this paper, we examine IP address allocation in cellular data networks, with emphasis on understanding the feasibility of IP-based geolocation techniques. We believe this is an important problem as smartphones connected to

cellular networks are increasingly being used to access Internet-based services. Of course, customized smartphone applications can use the built-in Global Position Systems (GPS) receiver to obtain accurate location information. However, in cases where a service is accessed through the phone's browser[1] or when GPS-based tracking is disabled (e.g., by the user because of privacy concerns), alternative geolocation techniques are necessary. The IP geolocation problem has not received much attention in the context of cellular data networks, and we fill this void by instrumenting a popular location-based iOS application to collect and subsequently analyze a dataset that has GPS-based location data for approximately 29,000 cellular network assigned IP addresses, obtained from several thousand individual smartphones spread across 50 countries.

This paper offers several contributions. First, we characterize the dataset and offer insights on the global deployments of cellular data networks. For instance, we find that NAT and other middleboxes are widely deployed in cellular networks worldwide. We also provide evidence of service differentiation, where a provider assigns publicly visible IP addresses to some users, while other users are behind NAT boxes. Second, we study whether or not geolocation databases provide good location estimates and show that the error is 200km or more in 50% of our measurements. Further, we observe some large errors, owing to mobile operator's implementation of roaming functionality. This can be expected to become a commonplace problem as roaming traffic charges drop. Finally, we provide an original dataset to the community, with an unprecedented number of ground truth measurements of IP to geolocation mapping for cellular data networks.

The remainder of this paper is organized as follows. Section 2 present an overview of related work. Our data collection method and a preliminary analysis of the dataset is present in Section 3. An analysis of the IP addresses observed in our dataset is presented in Section 4. Section 5 presents concluding remarks.

2 Related Work

The problem of geolocating hosts in networks has been widely studied [8]. Techniques range from measuring packet latencies to landmark nodes and then estimating their location relative to these nodes [6–8], applying machine learning to ground truth datasets [5], or using tabular storage of IP prefixes and associated locations ('GeoIP databases') [1, 2]. The accuracy of GeoIP databases has also been debated [9, 10]. For instance, Poese et al. [9] recently evaluated the accuracy of several GeoIP databases using ground truth information from several POP locations from one European wired ISP and found that while most GeoIP databases can claim accuracy at the country level, their databases are heavily biased towards few countries.

[1] The HTML5 Geolocation API [3] allows browsers to report a device's position. The source of location data is implementation-dependent, and can be obtained from GPS receivers, WiFi network location databases, or other means. It is still early days for this solution, and geolocation databases are likely to be a popular method for many reasons, including privacy concerns associated with fine-grained location tracking.

Data	Description
Unique ID	Per device, unique id (fully anonymised)
Timestamp	Time at server when measurement was recorded
Interface IP Address	IP address assigned to the Cellular Data interface
Observed IP Address	Device IP address, as observed at the application's server
Location	Latitude / Longitude coordinates
Horizontal Accuracy	Accuracy, in meters, of the location measurement

Table 1. Dataset details.

Closely related to our work are recent studies by Balakrishnan et al. [4], Xu et al. [12], and Wang et al. [11]. For mobile devices connected through 3G networks, Balakrishnan et al. [4] studied the accuracy of GeoIP databases, the client/server latencies, and the IP address 'stickiness'. Their study, while comprehensive, is based on three datasets with a maximum of about 100 devices, over a single mobile operator network in the US. Xu et al. [12] combined several data sources to discover cellular network infrastructure. Their work relied on server logs, DNS request logs, and publicly available routing updates to characterize four major US cellular carrier networks. Xu et al. evaluated the cellular network diameter, and demonstrate how this could affect content placement strategies. Wang et al. [11] characterized NAT, firewall, and other security policies deployed in more than 100 cellular IP networks.

We believe our work complements these recent efforts [4, 11, 12]. Our novel dataset has ground truth information on the location of mobile devices, and thus allows us to evaluate how well GeoIP databases may perform for IP addresses assigned by cellular networks. Further, our dataset provides an opportunity to study IP address assignment at a larger scale than that of previous studies, and across carriers in many different countries.

3 Dataset and Preliminary Analysis

3.1 Dataset

Use of third-party smartphone applications has exploded in recent years, owing to the phenomenal success of the 'App Store' model. These third-party applications present an unprecedented opportunity for crowd-sourcing network measurements from mobile networks. For this work, we partnered with the developer of a location-based iOS application[2] to add minimal instrumentation code such that the application's Internet-based server logs reported the device's local IP address. This reporting is only done when the device is using the 3G/GPRS interface for communication.

[2] The application is available only on Apple devices running the iOS operating system, and has been downloaded by 140,000 users from 50 countries, and is particularly popular in Germany and Australia.

The application developer provided us with processed data from their server logs. In particular, the raw dataset consists of 29,043 measurement points, collected from 11,230 unique smartphones between May and August 2011. The information available is detailed in Table 1.

This dataset may be obtained by contacting the authors. For privacy reasons, the released dataset will not provide the location data and instead provide the corresponding country and city-level information available from the Google reverse geocoding service. In addition, the released dataset will include the observed IP address but not the Interface IP address. Instead, we include a set of two boolean flags, to indicate respectively whether the device IP address was in the private IANA space, and whether it was different from the observed IP address. Finally the device id and horizontal accuracy are also removed. This transformation on the data improves the users privacy while providing the information required to confirm the key results of this paper, and develop many new findings.

3.2 Geographical Coverage

Before analyzing the collected data, we applied a few simple filtering rules. Note that the number of measurements from a particular device depends on the frequency with which the owner of the device interacts with the application. As we are not interested in recording multiple instances of identical information, for each smartphone we discard a measurement point only if all the following conditions are met, with respect to the previous measurement point: i) both the device and observed IP address are unchanged, ii) the distance between the measurements locations is less than 1km[3], and iii) the time elapsed since the previous measurement is less than 3 hours. Following this preprocessing, we are left with 27,328 measurements. Next, we applied the Google reverse-geocoding service to obtain city and country information from the GPS coordinates. We successfully looked up 26,566 dataset entries. The remainder of this paper focuses on this filtered dataset. In total, we have measurements from 1,924 cities in 50 different countries as summarized in Table 2 and illustrated in Figure 1.

Devices running iOS use a proprietary 'Assisted GPS' method to optimize the device location computation, using a combination of GPS data and a proprietary WiFi geo-database. The 3G iPhone's Assisted GPS typically has horizontal accuracy errors of 10-15 m [13]. The iOS application programmer can retrieve the accuracy level, in meters, associated with any GPS location measurement. This horizontal accuracy value was available for 97% of our measurement points, and these indicate that 78% of the GPS location information are accurate to 100m, and 93% are accurate to at least 1km.

[3] Condition (ii) captures mobility and uses 1km as the threshold since more than 90% of the measurements have horizontal accuracy of at least 1km.

Continent	Countries	# of Cities	Total Entries
Australia (2)	AU,NZ	166	18,211
Europe (26)	DE,FR,SE,AT,CH,GB,ES,IT,PL TR,LU,DK,BE,GR,NL,HU,RO RS,FI,CZ,HR,NO,IE,LI,PT,SK	1482	7,036
Asia (14)	TW,SG,JP,MY,CN,HK,KW KH,CY,OM,IN,AE,KR,LB	158	991
America (3)	US,CA, CR	104	282
Others (5)	MO,AR,CO,PR,ZA	14	46

Table 2. Reverse geocoding of measurement locations.

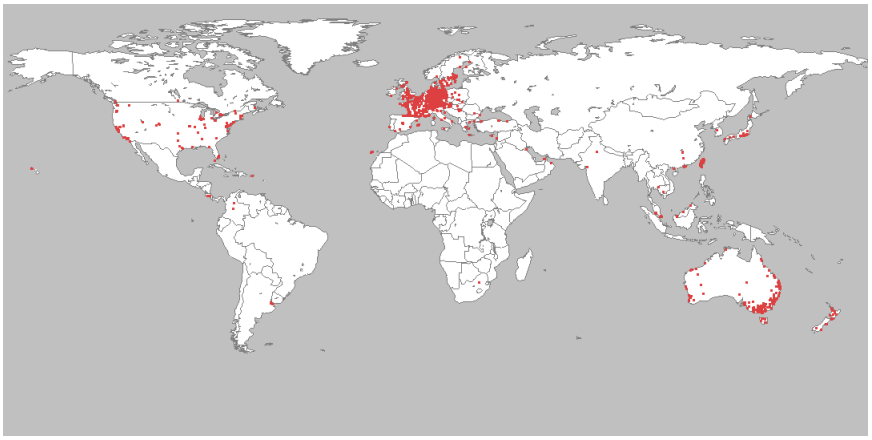


Fig. 1. All measurement locations.

3.3 Limitations

Our dataset constitutes a sample of smartphone locations worldwide, the IP address assigned by the cellular data network to these smartphones, and the IP address from which these devices are visible on the Internet. The main drawback of this application-driven measurement method is the spatial and temporal sampling bias introduced as the measurement occurrences are driven by: i) the adoption rate of the smartphone type/OS on which the application is available, ii) the application adoption rate and the spatial distribution of its adopters, and iii) the application usage rate and spatial pattern, which is dependent on the application's intended use. This dataset is, however, to our knowledge, the first of its kind to be available to the research community.

4 Cellular Networks: View from the IP level

4.1 Public IPs, Private IPs, and Middleboxes

With the number of Internet-enabled smartphones exploding, and the increased scarcity of available IPv4 address space, mobile operators are likely to rely on

Network Name	Country	total devices	# devices with only private IP	# devices with only public IP
OPTUSINTERNET-AU	AU	2039	11	1958
CUSTOMERS-DE	DE,IT,HR, FR,PT,NL	1337	1134	135
TELSTRAINTERNET42-AU	AU	1122	1119	0
VODAFONE	AU	1101	1029	59
H3GAIPNET	AU	789	783	1
DE-D2VODAFONE	DE,ES,NL,CH, FR,IT,DK,GR	702	692	8
VODAFONE-PACIFIC-AU	AU,NL	486	479	0
E-PLUS-MOBILES-BLOCK-6	DE	342	341	0
o2-Germany-NAT-Pool2-FRA	DE	300	299	0
o2-Germany-NAT-Pool1-DUS	DE,ES	283	282	0
o2-Germany-NAT-Pool1-BER	DE	265	264	0
DE-D2VODAFONE-20101118	DE	217	216	1
ORANGE-FR	FR	183	183	0
SFR-INFRA	FR,BE	163	163	0
EMOME-NET	TW	162	3	159

Table 3. IP allocation statistics for the top 15 networks in the dataset.

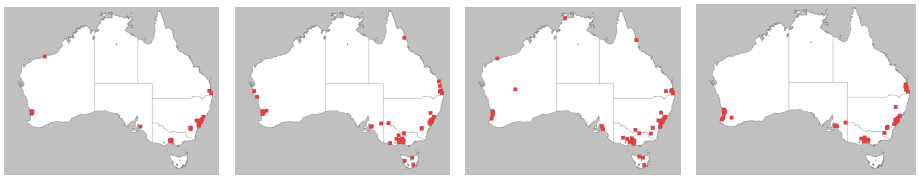
Network Address Translation (NAT) solutions. This section quantifies, for our dataset, the prevelance of public IP address assignment, NAT solution, and other middleboxes.

We observe 18,949 measurements, or roughly 70% of the measurements, where the smartphone's device interface is assigned a private IP address. Assignment of an address from IANA's reserved IP address space indicates the use of NAT solutions between the user's device and the application server. Further, we identify 478 instances where the phone's cellular interface address is assigned a public IP address but it does not match the observed IP address at the application server, thus indicating the presence of middleboxes between the device and the application server.

Table 3 illustrates the diversity of networks seen in our dataset. For each measurement point, we determine the network operator by querying the WHOIS service. The query uses the interface IP address if it is public or the observed IP address if the interface is assigned a private IP address. The table ranks networks based on the number of unique smartphones matched to a network. We notice that most operators use some form of NAT. Further, among these top 15 networks, we find several instances where a network assigns private IP addresses to some devices and public IP addresses to other devices, indicating service differentiation within operators: some devices benefit from publicly routable IP addresses, but most do not. We also find a few instances where a smartphone is assigned a private IP address at one point in time, and a public IP address at another point in time.

/24 IP block	# Countries	# Measurement	Country List
77.24.0	7	246	DE,FR,NL,DK,IT,ES,CH
80.187.96	4	174	DE,NL,IT,FR
193.247.250	4	88	FR,IT,NL,CH
80.187.107	3	303	DE,HR,PT
203.20.35	2	792	AU,NL
80.187.106	2	360	DE,IT
89.204.153	2	359	DE,ES
80.187.110	2	310	DE,FR
80.187.111	2	281	DE,FR
80.187.97	2	180	DE,IT

Table 4. /24 IP blocks with hosts in more than one country.



(a) 202.81.69.a (b) 58.163.175.b (c) 192.148.117.c (d) 58.163.175.d

Fig. 2. Dispersion of hosts around the top four mobile gateways in the dataset.

4.2 Spatial Allocation of IP Blocks

We investigated the geographical span of devices belonging to the top /24 IP subnets in the dataset. This indication can be useful when building GeoIP databases, especially when longest prefix matching strategies are used. We identified the top 10 /24 subnets that account for the most measurements from unique devices, and used Google's reverse-geocoding service to lookup the country location for each measurement in this set. Using a WHOIS service, we verified that all IPs in each /24 subnet does indeed belong to the same network provider. Table 4 summarizes our results, and illustrates that devices physically present in different countries may be assigned an address from the same IP block.

4.3 Spatial Coverage of Gateways

When a cellular network uses middleboxes, the application's server will receive connections from several distinct devices, all originating from a single IP address (which we refer to as the mobile gateway IP address). Here, we study the spatial dispersion of devices around mobile gateways, as observed in our dataset. This has implications on the accuracy of GeoIP databases as multiple, potentially far apart, devices have the same IP address from the point-of-view of Internet servers.

Figure 2 illustrates the dispersion of hosts attached to some example mobile gateway IP addresses[4], for one country (Australia). It is interesting to note

[4] The host number part of the IP addresses is truncated.

Observed IP	# Country	Country (# Measurement)
77.24.0.a	3	DE(28),IT(1),NL(1)
77.24.0.b	3	DE(21),ES(2),NL(1)
193.247.250.c	3	CH(2),FR(1),NL(1)
203.20.35.d	2	AU(532),NL(1)
77.24.0.e	2	DE(47),ES(1)
77.24.0.f	2	DE(34),CH(1)
77.24.0.g	2	DE(27),DK(1)
77.24.0.h	2	DE(24),FR(1)
202.175.20.i	2	MO(8),CN(3)
89.204.153.j	2	DE(8),ES(1)

Table 5. Top 10 observed gateway addresses with hosts in more than one country.

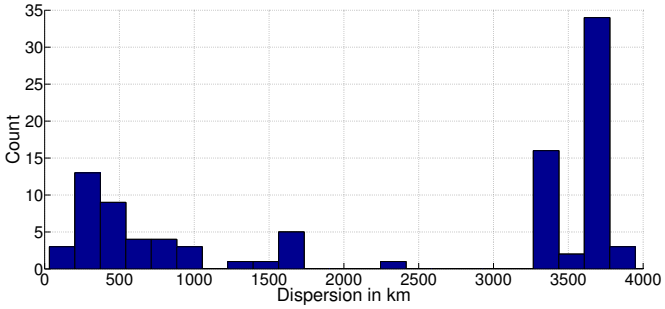


Fig. 3. Geographical dispersion of mobile hosts around the top 100 gateways.

that each gateway has hosts roughly in all major Australian cities. In addition, we found one device in the Netherland with the observed IP of 203.20.35.d, which is most likely a roaming user. Table 5 quantifies the spatial diversity for the top 10 gateways with hosts in more than one country, in our dataset. Our dataset suggests that mobile networks allocate IP addresses at a country-level granularity: mobile hosts exit the operator's network through a few gateways within the country, and these exit points may also be maintained while roaming.

We quantify the geographic spread of hosts served by a gateway by computing the maximum distance between any two hosts that are connected to the Internet through the same gateway. Figure 3 shows the histogram of the maximum dispersion values (in KM), for the top 100 gateways in our dataset. We notice that there are three clusters: one at approximately 500km, one at about 1500 km, and another at about 4000km. These clusters approximately correspond to the average inter-city, inter-state, and inter-country or inter-continental distances in our dataset. We also observed an outlier at 17,000km (not shown on the plot) which correspond to an Australian user roaming in the Netherlands.

4.4 Accuracy of IP Geolocation Databases

We also tested the ability of GeoIP databases to return host location based on IP addresses seen by the application's server. For this analysis, we use two commercial GeoIP databases, namely MaxMind [2] and IPinfoDB [1], and compute the

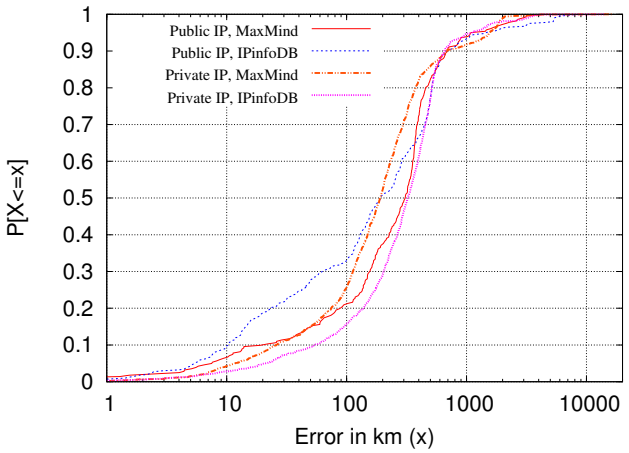


Fig. 4. Geolocation error when using GeoIP databases.

error as the distance between the geographical location returned by the GeoIP database and our measurement location. As previously mentioned, application-level measurements introduce sampling bias: as each measurement occurrence depends on a user starting the application and the user's current position, more popular areas or areas where the service is more popular will have more measurement points. To address this spatial sampling bias, we normalize the error at the city scale, by computing the average error for each city (identified using Google's reverse geocoding service).

Figure 4 shows the distribution of the computed errors, for the GeoIP databases considered, with results presented separately for public and private IP addresses (Note that for devices with private IPs we use their gateway address as visible to the server on the Internet.) For our dataset, depending on the database used, errors of 100km or more are observed in at least 70% of the measurements although 90% of the errors are under 1000km. The errors are typically larger for devices assigned private IP addresses. This is not surprising as we have previously noted that one mobile gateway could potentially cover an entire country, including countries as large as Australia.

5 Concluding Remarks

We studied cellular data networks from the point-of-view of IP clients, covering both spatial and IP-layer aspects. Our work is based upon a comprehensive dataset of several thousand mobile device locations and IP addresses. Our dataset suggests that mobile operators worldwide are using some form of NAT or middlebox. This has implications for application designers (e.g., difficulty of implementing peer-to-peer communication, performance implications). As hosts behind NATs appear from a few IP addresses per country, we shown how these IP addresses can cover hosts physically present in entire countries, across international borders, and even continents. We also evaluated the accuracy of GeoIP

database in the mobile domain, and found that, for our data, the distance error between the GeoIP database determined location and the GPS determined location is at least 100km for approximately 70% of our measurements, with a few errors being substantially larger.

Acknowledgements

This work was supported by the Commonwealth of Australia under the Australia-India Strategic Research Fund.

References

1. IPInfoDB. http://ipinfodb.com.
2. Geolocation and Online Fraud Prevention from MaxMind. http://www.maxmind.com/, 2011. [Online; accessed 14 Sept-2011].
3. Geolocation API specification. http://www.w3.org/TR/geolocation-API/, 2011. [Online; accessed 14 Sept-2011].
4. M. Balakrishnan, I. Mohomed, and V. Ramasubramanian. Where's that Phone?: Geolocating IP Addresses on 3G Networks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference*, pages 294–300, Chicago, IL, November 2009.
5. B. Eriksson, P. Barford, J. Sommers, and R. Nowak. A Learning-based Approach for IP Geolocation. In *Proceedings of Passive and Active Measurement Conference*, pages 171–180, Zurich, Switzerland, April 2010.
6. E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP Geolocation using Delay and Topology Measurements. In *Proceedings of ACM SIGCOMM Internet Measurement Conference*, pages 71–84, Rio de Janeiro, Brazil, October 2006.
7. S. Laki, P. Mátray, P. Hága, I. Csabai, and G. Vattay. A Model-based Approach for Improving Router Geolocation. *Computer Networks*, 54(9):1490–1501, 2010.
8. V. Padmanabhan and L. Subramanian. An Investigation of Geographic Mapping Techniques for Internet Hosts. In *Proceedings of ACM SIGCOMM*, pages 173–185, San Diego, CA, August 2001.
9. I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, 41(2):53–56, April 2011.
10. Y. Shavitt and N. Zilberman. A Geolocation Databases Study. *IEEE Journal on Selected Areas in Communications*, 19(10):2044–2056, December 2011.
11. Z. Wang, Z. Qian, Q. Xu, Z. Mao, and M. Zhang. An Untold Story of Middleboxes in Cellular Networks. In *Proceedings of ACM SIGCOMM*, pages 374–385, Toronto, ON, August 2011.
12. Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao. Cellular Data Network Infrastructure Characterization and Implication on Mobile Content Placement. In *Proceedings of ACM SIGMETRICS*, pages 317–328, San Jose, CA, June 2011.
13. P. A. Zandbergen. Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, 13(S1):5–25, June 2009.