GPS-Based Geolocation of Consumer IP Addresses

James Saxon(✉) and Nick Feamster

University of Chicago, Chicago, IL 60637, USA
{jsaxon,feamster}@uchicago.edu

Abstract. This paper uses two commercial datasets of IP addresses from smartphones, geolocated through the Global Positioning System (GPS), to characterize the geography of IP addresses from mobile and broadband ISPs. Datasets that geolocate IP addresses based on GPS offer superlative accuracy and precision for IP geolocation and thus provide an unprecedented opportunity to understand both the accuracy of existing geolocation databases as well as other properties of IP addresses, such as mobility and churn. We focus our analysis on three large cities in the United States.

After evaluating the accuracy of existing geolocation databases, we analyze the circumstances under which IP geolocation databases may be more or less accurate. Within our sample, we find that geolocation databases are more accurate on fixed-line than mobile networks, that IP addresses on university networks can be more accurately located than those from consumer or business networks, and that often the paid versions of these databases are not significantly more accurate than the free versions. Addresses on /24 subnets that are geographically concentrated are geolocated more accurately. We then characterize how quickly /24 subnets associated with fixed-line networks change geographic locations, and how long residential broadband ISP subscribers retain individual IP addresses. We find, generally, that most IP address assignments are stable over two months, although stability does vary across ISPs. Finally, we evaluate the suitability of existing IP geolocation databases for understanding Internet access and performance in human populations within specific geographies and demographics. Although the median accuracy of IP geolocation is better than 3 km in some contexts – fixed-line connections in New York City, for instance – we conclude that relying on IP geolocation databases to understand Internet access in densely populated regions such as cities is premature.

1 Introduction

IP geolocation is a longstanding problem in computer networking, with both an active academic research and a wide array of commercial solutions and applications. IP geolocation is used for a variety of purposes, including mapping clients to nearby content delivery network (CDN) replicas, personalization of search results and advertising, and customization of content (e.g., weather or language

O. Hohlfeld et al. (Eds.): PAM 2022, LNCS 13210, pp. 122–151, 2022.
https://doi.org/10.1007/978-3-030-98785-5_6

localization). In a legal context, IP geolocation is used for digital rights management (e.g., geographic licensing restrictions), compliance with the laws and regulations of a region or country (e.g., gambling, sales taxes, privacy regulations), and to assist with law enforcement (e.g., determining jurisdictions or collecting evidence). In security contexts, government and commercial entities use it for counter-terrorism, attack attribution, monitoring access to private networks, and detecting potential fraud. It facilitates operations and site reliability (e.g., monitoring packet loss from a location), and informs infrastructure investments by both industry and policymakers [2,9,12,19,25,29]. Computer science researchers also use IP geolocation to study the properties and evolution of the network itself, such as the structure and graph parameters of networks [6,28].

Increasingly, IP geolocation is being used to address various problems in *policy and social science* that entail drawing inferences about various demographics and geographies based on inferred locations of IP addresses. Social scientists have noted the potential to use "big data" as a lens on human behaviors and interactions [18], and as modern society is increasingly mediated through the Internet, many of our interactions are associated with IP addresses. Server logs and speed measurements, for instance, show who accesses resources and the quality of their connections. This allows aggregate statistics or time trends. But associating these behaviors and network conditions with human populations ultimately requires a way to map IP addresses to physical locations. A natural approach would be to use IP geolocation with census tract-scale precision to link IP addresses to physical locations. In this paper, we leverage reference locations of unprecedented geographic precision to evaluate whether free and paid IP geolocation databases can achieve this level of accuracy in large cities in the United States. We also extend past work by analyzing the determinants of IP geolocation accuracy – the IP addresses for which geolocation is or is not reliable. We then interpret these findings with a view towards social research, describing *who* gets lost from a naïve reliance on IP geolocation, and what the consequences might be for academic or policy analysis.

The accuracy of IP geolocation databases has practical implications for the answers to a wide range of social and public policy questions. One area of particular timeliness is that of the so-called "digital divide." Calls for digital equity and inclusion, already urgent, have reached a fever pitch during the COVID-19 pandemic. Prominent studies of broadband performance from Microsoft and M-Lab rely on IP geolocation to associate Internet throughput and latencies with zip codes [17,24]. Ganelin and Chuang studied whether or not geolocation databases could reliably indicate socioeconomic status of MOOC registrants with known physical addresses. That study ultimately concluded, as will we, that answering such questions based on existing IP geolocation databases is premature [7].

We revisit this problem now, due both to its practical implications, and thanks to the availability of two highly-accurate and large-scale groundtruth datasets of GPS-located IP addresses. These datasets, from Unacast and Ookla® Speedtest Intelligence®, afford us a view of consumer behaviors on both fixed-line and mobile networks, that is markedly different from the geolocation targets used in past work.

Table 1. Main findings, with pointers to sections.

§	Main findings
4.1	GPS reports are a credible groundtruth of IP locations
4.2	MaxMind's GeoIP2 service provides the lowest median error among tested services and cities: 2.62 km on fixed-line addresses in NYC
5.1	IP geolocation performs better on fixed-line consumer networks and universities, and worse on mobile broadband and businesses
5.2	The physical size of subnets is correlated with the accuracy with which they are IP geolocated
5.3	On the two-month time-scale, the median fixed-line /24 IPv4 subnet in US cities moves less than 1 km
5.4	Churn of individual IP addresses on fixed-line networks in major US cities takes months
6	Access modalities – mobiled vs fixed – differ between demographic groups. Even on fixed-line networks, relying on IP geolocation to identify neighborhoods would lead to biased results

Table 1 lists our main findings. The rest of this paper is organized as follows. Section 2 discusses related work in IP geolocation, both in research and in commercial product offerings. Section 3 describes the datasets that we use for the analysis in this paper. Section 4 evaluates the quality of the datasets that we are using, in particular exploring the suitability of using GPS data as a "ground truth" for evaluating IP geolocation databases. Section 5 presents the result of our study, including findings about the circumstances under which IP geolocation is more or less accurate. In Sect. 6, we interpret and extend our results in the context of research on human populations and privacy. We conclude in Sect. 7.

2 Related Work

Past work on IP geolocation generally takes three approaches, as outlined by Padmanabhan and Subramanian [23]. Their IP Geolocation work, IP2Geo, compared the complementary strengths of active latency measurements (GeoPing), active traceroutes paired with DNS hints (GeoTrack), and static databases of outside information (GeoCluster). Each of these approaches has evolved. Padmanabhan and Subramanian concluded that database-driven methods held the greatest promise. Commercial products have accordingly built databases with proprietary methods that include registry information, outside data, and active methods. On the other hand, academic work has tended to focus on active and DNS-based measurements.

IP Geolocation Methods. Starting with DNS, Spring et al. developed techniques in their Rocketfuel project to map infrastructure (i.e., routers) to physical locations. A significant contribution was to optimize traceroute targets to minimize redundancy and ensure that each path will traverse its target ISP [28],

although their use of the DNS to geolocate routers was pioneering at the time. Their subsequent approach to DNS hint identification was largely manual – "browsing through the list of router names" – but the resultant `undns` tool has proven influential and enduring. Freedman et al. extended `undns`' coverage [6]. These projects were driven by questions about properties of the network, specifically the topology of large ISPs and the efficiency of block assignments in BGP routing tables. More recently, Dan et al. [2] attempted to enumerate all possible DNS city name hints and finalize location decisions with machine learning. Like IP2Geo, the authors relied on a large dataset from Microsoft for their ground truth, although the ground truth data was from Bing instead of Hotmail.

In the latency-based space, Gueye et al. [12] and Katz-Bassett et al. [15] introduced constraint-based geolocation (CBG) and topology-based geolocation (TBG). CBG is essentially the intersection of several latency-derived distance buffers, while TPG also localizes intermediate hosts so that targets can be constrained by their relation to passive landmarks rather than just active probes. Subsequently, Octant incorporated both positive *and negative* constraints (the IP address is *not* within a certain radius) [30].

In addition to this "geometric" approach are several statistical strategies. Eriksson and colleagues, developed first a Bayesian approach and then a likelihood-driven choice among possibilities with the CBG-derived regions [3,4]. Other work presents strategies using kernel density and maximum likelihood estimation [1,31]. It is also possible to constrain location from the covariance matrix of latency measurements with locations.

Notable in Eriksson's Bayesian work is the insight that outside information can help constrain or inform geolocation. They used population as a measure of places' importance, as have later researchers [2]. Other forms of information help as well. In trace-based work reminiscent of TPG, Wang et al. performed extensive webscraping and analysis to identify and confirm businesses with locally-hosted sites that they could "enlist" as passive landmarks. They used those landmarks to identify the locations of routers near the geolocation target [29].

Scalability has long been a limitation of active measurements. Since locations are most-constrained by the closest locations, Hu et al. developed methods to prioritize measurements from nearby hosts, effectively by localizing avatars from subnets [13]. Alternatively, Li et al. "flip" the standard infrastructure of active geolocation with GeoGet: the targets to be localized measure the latency themselves, through javascript, rather than generating pings through an API [19]. This reduces the number of servers and traffic required, and it is also helpful since clients' devices or networks may fail to respond to pings or complete traceroutes.

Evaluating Commercial Services. These advances notwithstanding, commercial geolocation tends to be implemented through databases, which are inexpensive to distribute and can aggregate historical observations across many sources. The leading services—MaxMind, IP2Location, Akamai, or NetAcuity—all use proprietary methods. A number of papers assess the performance of these databases, comparing with the preceding active methods [11], points-of-presence

paired with routing tables from a large ISP [25], DNS lookups paired with ground truth rules from domain operators [8], from RIPE ATLAS built-in measurements, or PlanetLab nodes, or against each other, sometimes with a majority logic applied. The databases are themselves often taken as the ground truth for latency-based measurements often with a sort of majority logic. Shavitt and Zilberman employ that strategy in evaluating the databases themselves, but also focus on *consistency* among addresses determined to share a point-of-presence, based on an earlier algorithm [5,27]. Similarly, Huffaker et al. assess the agreement of country determinations and distances from a centroid, from majority votes (supplemented by PlanetLab ground-truth and limited round-trip time measurements) [14].

On the whole, both the formal literature and "popular wisdom" paint a fairly pessimistic picture of geolocation performance. Research studies from about ten years ago assessed median accuracy of these services at 25 km in Western Europe and 100 km in the United States. On the commercial side, Poese et al. quote median accuracies between tens and hundreds of kilometers for MaxMind and IP2Location [25]. Other early works present distributions with ranges between hundreds or thousands of kilometers [27]. Gharaibeh et al. present results for routers in particular, with median accuracies between 10 km for NetAcuity and 1,000 km for IP2Location, on either extreme of the free and paid versions of MaxMind. More recently, Dan et al. presented medians between 10 and 30 km, depending on the sample and service. [2] They present results in 10 km bins and do not differentiate performance at the very bottom of the range.

Studies of How Internet Infrastructure Affects Geolocation Accuracy. A persistent though somewhat more subtle current of the literature has explored the physical structure of the Internet and its relation to geolocation accuracy. Padmanabhan and Subramanian anticipated the interplay between network infrastructure and geolocation accuracy in 2001 [23]. They noted the impact of the geographical concentration of AOL's login nodes on accuracy, and showed that clusters of addresses that were physically larger were associated with poorer performance for the GeoCluster (database) method. This point was echoed in 2007 by Gueye et al. [11] Similarly, Freedman et al. measured the physical scale of autonomous systems. Later, Gharaibeh et al. probed the common assumption of databases that /24 subnets are co-located [11] Those papers show that systems, subnets, and IP prefixes advertised by the Border Gateway Protocol (BGP) can span large physical distances. In this paper, we seek to extend this work, aiming to identify the circumstances when they are large or small. Huffaker et al. characterized accuracy according to carriers' network role; we extend that line of exploration in this research, exploring how accuracy varies between commercial ISPs, large companies, and universities. We categorize addresses by "Doing-Business As" names reported in IP address registries; to our knowledge, such a characterization is unprecedented, at least in the current era where mobile devices are significantly more prevalent than they were a decade ago.

In addition to work on IP address *locations*, our data also shed light on the persistence of dynamically assigned IP addresses, itself an active area of analysis. Recent works have used RIPE Atlas probes [16,21], javascript-based user

monitoring by a large CDN [22], and browser extensions [20] to study address retention times. Times range from nearly-ephemeral on mobile networks, to many months for fixed-line connections in the North America. The retention times we observe are broadly consistent with previous findings for North America.

Finally, our project is informed by recent work on Carrier-Grade Network Address Translation (CG-NAT). CG-NATs are increasingly common across all ISPs, but almost ubiquitous on mobile carriers [26]. Since we geolocate public IP addresses, it stands to reason that the geolocation accuracy of devices behind a CG-NAT cannot be more precise than the basic spatial scale over which a public address is used. Nevertheless, the physical extent of CG-NATs' structures have not been studied. Public IP addresses could map to limited geographic locations like antennas, or to larger ones like cities.

How this Paper Extends Past Work. Past work that evaluates IP geolocation accuracy has tended to rely either on active measurements of somewhat coarse precision, or on a fairly consistent set of (unrepresentative) benchmarks: specifically, PlanetLab sites and university clusters. The dataset we rely on for this paper of course has its own peculiarities—it is a non-random sample of mobile devices— but this view from the access network, including mobile devices, is critical and distinctive from past studies. It is a large sample, indicative of realistic consumer geolocation targets in major cities in the United States. The Global Positioning System (GPS) has long served as a counterpoint to IP geolocation, both as a benchmark of accuracy and as an analog in multilateration. Historically, its deployment and use for Internet measurement felt impossibly far off [3,4,15], but the future has now arrived.

This paper complements and extends previous work as a result of its large sample of consumer smartphone locations on diverse networks. The primary dataset was provided by Unacast; we confirm our basic findings with a smaller, Chicago-only sample of GPS-located Speedtest® data from Ookla®. Similar datasets are readily available for commercial applications and academic research. We exploit this sample to understand how IP geolocation accuracy varies by geography, carrier, mode of access, and other factors. In contrast to previous work, which has tended to question the overall reliability of geolocation even at country-level accuracy, we find that it works fairly well in predictable and well-defined contexts. Nevertheless, the imperfect accuracy and context-specific performance still currently constrain the applicability of IP geolocation for studying Internet access by human populations.

3 The Data

This paper relies on two commercial datasets with GPS-tagged IP addresses to analyze the geography of consumer IP addresses. We also evaluate and analyze the performance of databases for IP geolocation from two popular, commercial services: IP2Location and MaxMind from the same time periods. Table 2 lists the datasets that we use, and Fig. 1 illustrates how these datasets are joined and augmented in our analysis.

Table 2. Data sources: geographic and temporal coverage, and data volumes (for GPS data only).

	Geography	Date	Location reports
Unacast Clusters	NYC, Chi., Phl	Aug-Oct 2020	248M
	+ 40 mi buffer	Apr 2021	9.5M
Speedtest Intelligence	Chicago Region	2020	4M
MaxMind Free	Global	Aug 2020	–
	Global	Apr 2021	–
MaxMind Paid	North America	Apr 2021	–
IP2Location Free	Global	Aug 2020	–
	Global	Apr 2021	–
IP2Location Paid	Global	Apr 2021	–

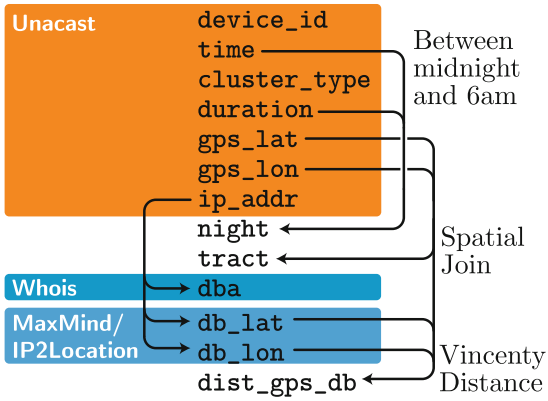

Fig. 1. Simplified illustration of the data augmentation process, for Unacast data. The fundamental data consist of device identifiers, times, locations, and IP addresses. Clusters (see text) are also labelled by type, for instance, `TRAVEL` or `LONG_AREA_DWELL`. The time and duration are used to construct a flag for night-time clusters. The IP address is used with the ARIN whois resource to construct Doing Business As (DBA) names, and database-defined locations are retrieved from up to four databases by MaxMind and IP2Location. Vincenty distances are calculated between database and GPS locations

The GPS data were delivered anonymized and remain so. The data were collected in accordance with local laws and opt-out policies (GDPR), and analyzed with approval from our university's Institutional Review Board (IRB). The IRB approved analysis of reconstructed "home locations" for earlier work, but emphasized the sensitivity of doing so. For that reason, we avoided geographic analysis of individual devices in this project, and proxied "residence" simply as activities recorded at night.

3.1 Unacast GPS Smartphone Locations

The primary dataset used for the analysis is from Unacast, a location intelligence firm. This dataset contains GPS locations reported by mobile devices, along with timestamps and unique, anonymous identifiers. Unacast aggregates multiple location data streams from other firms; they perform extensive data validation, de-duplication, and processing on those streams. The exact applications that generate locations are not provided. The share of data reporting IANA reserved or private addresses is low, at 0.5%, and the share of addresses associated with foreign Internet registries totals just 0.2% (mostly RIPE, breakdown shown in the Appendix). The traffic observed in the Unacast dataset is overwhelmingly IPv4, at 99.6%.

We were provided with data for three major cities in the United States: New York, Chicago, and Philadelphia. Data were drawn from a 40 mile buffer of each city's boundaries; this large buffer encompasses both urban and rural populations. Two samples were provided in time. The first was from August–October 2020. A second, shorter period from April 2021 was provided to align with licenses for paid geolocation databases, allowing us to evaluate the accuracy of those services. As discussed below, the IP address from which a physical location is reported is recorded for about half of clusters in the 2020 sample, although this falls to just 15% in the 2021 sample. Data are used only when they contain an IP address, and the full dataset thus offers IP addresses recorded at over 248 million locations. Of course, individual IP addresses may be reported many times.

The data also report an estimate of the GPS-based location accuracy; the median reported accuracy is 17 meters on the 2020 sample and 11 meters on the 2021 sample. A small fraction of data (1.7%) are recorded with four or fewer decimal points of coordinate precision, corresponding to a physical distance of about 10 m. We exclude these data from subsequent analyses, along with location reports with estimated accuracy worse than 50 m. We also exclude the small fraction of addresses associated with private IP ranges and foreign NICs. These requirements do change the "Universe" of data included in the analyses that follow, and may impact CDFs.

Location Clustering and Classifications. Each line of data represents a *cluster* of location reports, called *bumps*. Clusters are built by combining bumps from an individual device that are close in both time and space, using Unacast's proprietary algorithm. That algorithm uses machine learning to account for variation in physical scale among locations: a mall is larger than a coffee shop or a home. Clusters are labelled according to their durations, which are also reported. Locations recorded during movement are labelled as `TRAVEL`. See the Appendix for a listing of cluster frequencies. This clustering reduces the data volume by a factor of 20 while retaining most of the information. Just as important, Unacast's data licenses with *its* suppliers often preclude re-licensing the raw, un-clustered data.

The clustering entails some subtlety: a single physical location and IP address is reported per cluster, and thus the centroid of a `TRAVEL` cluster may not exactly

coincide with the moment that the reported IP address was used. Indeed, the physical location of a consumer IP address is often not fixed; for instance, consumers can roam freely through their home while connected to their Wi-Fi. In practice, individual IP addresses are recorded at many physical locations—and these locations may be close or distant from each other.

Flagging Night-time Activity. We augment the provided data in several ways, illustrated in Fig. 1. As a means of selecting residential location reports, we flag clusters generated at night. Night-time clusters are those for which the period between the first and last bumps extends into the hours between midnight to 6am of any day. These clusters represent just 4.7% of clusters but 26% of bumps. Only 18% of devices have at least one night-time cluster, but those devices generate the vast majority of the data: 80% of clusters and 88% of bumps. In short, weighted by data volume, most devices have observations at times when they can reasonably be assumed to be at home. For the set of devices with night-time clusters, the ratio of devices to the population of the study region is about one device for every 20 people.

Identifying ISPs. To investigate the determinants of geolocation accuracy, we also identify ISPs. Each address is associated with its /24 subnet, whose organization is retrieved from the ARIN whois registry, on September 1 2020, or April 25 2021. If the prefix size of the associated CIDR block exceeds 24 on IPv4 or 48 on IPv6, we follow whois' link to the "parent" network. This strategy is similar in intent to an ASN lookup, and we include an ASN-based breakdown of ISPs in the Appendix. The whois look-up differs in practice primarily in superior coverage of the Department of Defense NIC and wireless carriers (AT&T and T-Mobile), especially for the RouteViews databases from August 2020. Further, the ASN lookup also "fractures" organizations like small city governments or businesses from their providers. We associate large and common organizations with standardized "Doing Business As" (DBA) names, taking particular care to capture the major ISPs in each market (Comcast, Charter, etc.). We separate AT&T's and Verizon's mobile broadband from their fixed offerings based on the words "Mobility" or "Wireless" in the organization name. This may not be a perfect division: "Verizon Business" and "AT&T Services" may include mobile offerings, but examining the ASN tables suggests this is not their primary use. It is worth noting that the sample is dominated by locations recorded while connected through mobile providers: there are ten times as many locations on AT&T mobile than AT&T fixed-line services, and more than five times as many on Verizon mobile than Verizon fixed-line. However, as we will separate addresses by ISP, this sample volume effect is largely "partitioned out." Ultimately, each address is associated with a single DBA name for analysis.

These procedures also identify large companies and institutions, in particular, universities. We flag addresses from universities with at least ten thousand students, and Fortune 100 companies. University clusters are "classic" targets for academic work on geolocation, since they have meaningful and well-known locations, but they are not representative of the consumer space. We exclude ISPs, including Google, from the Fortune 100 set. We tabulate IANA special use

and non-ARIN addresses, as checks on the underlying data, but exclude these from subsequent analysis.

3.2 Geolocated Ookla Speedtest Data

In addition to the data from Unacast, we have obtained Speedtest data from Ookla. The data are for tests performed on smartphones, again with locations from GPS. This dataset is substantially smaller, and is limited in geographic extent to the counties surrounding Chicago. We appeal to these data as a cross-check of the Unacast data that, though more voluminous, were not designed for this work.

We have received over 4 million individual Speedtest measurements for 2020, though only 270 thousand match the period of the study (August 2020). Unlike Unacast data, each location comes from a single moment in time (it is not a cluster). On the other hand, the Speedtest data include only the first three bytes of the IP address, due to privacy restrictions. We rely on Ookla's coding of Internet Service Providers.

3.3 Geolocation Databases and Distances

We obtain the free versions of the MaxMind and IP2Location databases, for August 1, 2020. We also acquire both the free and paid versions of these databases from April 26, 2021. The NetAcuity and Akamai geolocation services, which are much more expensive, are not included in this work. Using these databases, we geolocate IP addresses from the GPS sample. Per the license, this is done only for the months of GPS data matching the databases (August 2020 and April 2021). A recent review showed that MaxMind is by far the most-used database in the academic literature. It also found that databases change non-negligibly over short periods and emphasized that precision with respect to dates is imperative [10].

We then measure the Vincenty distance (on the ellipsoid of Earth) from each IP-geolocated point to the location recorded by the GPS-enabled device. For most of what follows, we take the centroids of the GPS clusters as the "ground truth" and call the entire distance the "accuracy" or "error." Since the database providers acknowledge their limited resolution and in certain cases quantify it accurately, this language is perhaps unfair: it is different for a database to acknowledge a location as unknown or indeterminate (as in reserved, private addresses) than to be "wrong" about the location. Moreover, the GPS data themselves do have some limitations, noted below. Semantics aside, the balance of this work tabulates distances with respect to the ground truth and seeks to explain their heterogeneity.

4 Evaluating Data Quality

Before coming to questions about the properties of consumer IP addresses, we analyze the quality of our data. We first explore the consistency of the GPS-

based location data we obtain from Unacast and Ookla by comparing the data against each other, with respect to geolocation databases.

4.1 Are GPS Data a Credible Ground Truth of IP Address Locations?

The accuracy of IP geolocation is central to Unacast's core business, and the company dedicates enormous resources to validating and maintaining their incoming data streams. While GPS data from smartphones is generally understood to be accurate, datasets from smartphone-based services do often incorporate additional data to assist with locating devices in circumstances where GPS does not work (e.g., indoors). Thus, while we expect these GPS-based datasets to be reasonably accurate in general, it behooves us to explore the quality of these datasets before proceeding with other questions. Since we aim to use these datasets as "ground truth", this analysis may seem a bit circular. Our strategy is to compare the *consistency* of IP geolocation results for different GPS contexts and across independent GPS samples (Unacast and Ookla). Of course, this analysis does not exclude the possibility of systematic errors arising in *both* GPS datasets, or across all datasets, but given the lack of further ground truths, we are left with consistency checks.

Direct cross-checks of IP addresses' locations between the two samples are not possible, because the Ookla data report /24 subnets rather than unique addresses. Many IP addresses are recorded at multiple physical locations, and in general a different set of addresses may be reported per subnet, in the two samples. Notwithstanding, the distance *can* be calculated between the medioids (the median of the x and y directions) observed for a subnet, in the two datasets. We do this for fixed-line providers in Chicago, on subnets with at least 10 distinct addresses and 10 distinct devices in the Unacast sample. If we weight subnets by the geometric mean of the number of observations in the two samples, the distance between their medioids is less than 2.5 km for 58% of subnets and less than 5 km for 83% of subnets.

Evaluating Cluster Types. The correspondence between GPS coordinates and the physical location of its IP address may not be perfect. For example, we expect that the clustering procedures could affect the "compatibility" of the IP address and GPS location. Further, if a GPS location is recorded when no network is available, it may be subsequently *reported* at a different physical location where an IP address can be obtained. We would expect these effects to be most severe for `TRAVEL` clusters, as previously discussed. The flip side of this argument is that navigation applications are more likely to be active during `TRAVEL`. These apps record location more frequently, which could *improve* accuracy.

To evaluate the effects of imperfect knowledge of locations, stemming from these effects, we contrast `TRAVEL` clusters with others. We will show below that geolocation performance differs by network. Obviously, it is easier to "travel" when connected to a mobile than fixed-line network. We therefore focus this

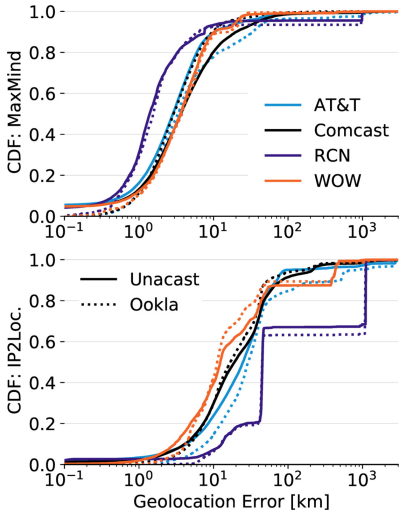

Fig. 2. Geolocation error of GPS location targets in Chicago, on both Unacast and Ookla Speedtest Intelligence® data, (Based on the authors' analysis of Ookla® Speedtest Intelligence® data for August 2020 in Chicago. Ookla trademarks used under license and reprinted with permission.) using the free versions of the MaxMind and IP2Location databases for August 2020.

check on a single, mobile network: AT&T Mobility. We do observe that accuracy is worse for travel than non-travel data, but the difference at the median is only about 2.5%, for either IP2Location or MaxMind. As can be seen in the Appendix, the cumulative distribution functions for travel and non-travel clusters are fairly close across their entire domain.

Analysis of Independent Samples. To further validate the GPS data, we contrast data from Unacast with Ookla, for fixed-line broadband ISPs, in Chicago and August 2020, where both datasets are available and aligned with the free versions of the geolocation databases. Figure 2 shows these results: the CDF of location reports as a function of geolocation accuracy. MaxMind performs somewhat better on Comcast addresses from the Ookla dataset than the Unacast data, and somewhat worse on AT&T; RCN and WOW! are very consistent. Discrepancies are somewhat larger on IP2Location as is comparative performance by the two databases.

One notable feature in the 2020 Unacast dataset is a small but non-negligible share of the data with IP geolocation "error" *very* close to zero. Depending on the ISP, that share is 4–5% of the fixed-line locations on MaxMind and 1–2% of those on IP2Location. On close inspection, these appear to be locations reported by applications *relying on the IP Geolocation services themselves*, rather than true GPS coordinates. For example, these ultra-"accurate" locations are not at residences, as one might expect for fixed-line ISPs, but in parks, as is MaxMind's practice for default locations [16,20]. The share of "too-close" locations is smaller

Table 3. Quantiles of accuracy in kilometers, for each database and city.

Quantiles	New York				Chicago				Philadelphia			
	MaxMind		IP2Loc.		MaxMind		IP2Loc.		MaxMind		IP2Loc.	
	Paid	Free	Paid	Free	Paid	Free	Paid	Free	Paid	Free	Paid	Free
0.10	0.7	0.8	3.0	3.0	1.0	1.0	4.6	4.7	1.1	1.2	4.2	4.2
0.25	1.4	1.5	6.1	6.1	1.8	1.9	9.8	9.9	2.1	2.3	9.0	9.0
0.50	2.6	2.8	12.0	12.1	3.3	3.6	24.0	24.3	4.0	4.3	20.9	21.0
0.75	5.0	5.5	30.1	30.5	6.4	7.1	45.6	45.7	7.6	8.2	39.6	39.7
0.90	9.7	11.0	61.8	63.0	13.0	16.1	196.4	202.9	13.5	15.0	78.3	78.8

on the 2021 clusters; however, the IP address field is populated for a lower share of those data.

However, the basic features of Fig. 2 are consistent in the completely separate sample from Ookla, which does not exhibit this feature.

4.2 Which Database Provides the Lowest Error in Location?

The practical question is which database to use, and how well it should be expected to perform. This analysis, uniquely, is performed using the April 2021 sample from Unacast, for which the paid geolocation databases were licensed. Since Sect. 5.1 will show that geolocation on mobile broadband is very poor, this analysis focusses on fixed-line broadband.

The short answer is that MaxMind's paid database, GeoIP2, provides the best accuracy, in terms of geolocation error on all quantiles. The traditional way of reporting this is the median error, which is 2.62 km in New York City, 3.31 km in Chicago, and 4.02 km in Philadelphia. Other quantiles and the other three databases are shown in Table 3. Figure 3 shows the distribution of distances by city and database. We use "city" to refer to the city itself along with the 40-mile buffer around it. Because the distance from Staten Island to North Philadelphia is only 46 miles, some data are included in the curves for both New York and Philadelphia.

Although the paid databases are more accurate in each city and at every quantile, the relative improvements in accuracy are modest. An important limitation of this particular study is our focus on urban areas in the United States. In particular, we do not test accuracy of these databases outside of major metro areas, and global or national performance may of course be different. Nonetheless, it would be possible to perform the analysis we have presented in this section for other datasets, if and when they are made available.

5 The Geography of Consumer Subnets

We now turn from an initial assessment of the dataset and databases, to measurements of the geography of the underlying networks.

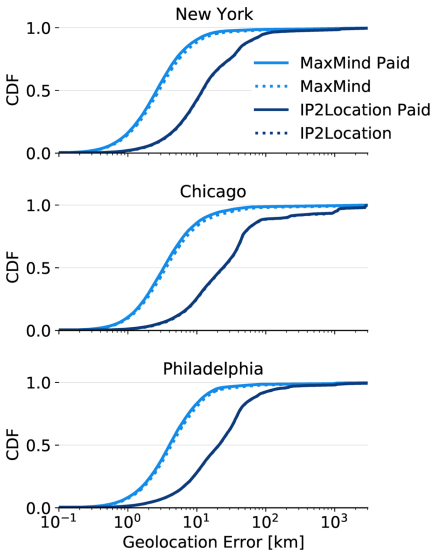

Fig. 3. Cumulative distribution function by geolocation database and city. Colors reference databases, and line styles denote paid and free versions. (Color figure online)

5.1 Under What Circumstances are IP Geolocation Databases Accurate?

The basic results of Sect. 4.2 mask extreme but unsurprising heterogeneity. Figure 3 already shows that geolocation performs better in New York than Chicago, and better in Chicago than Philadelphia. But the largest source of heterogeneity stems from providers, which deploy different physical infrastructures (and serve different cities). This entire section relies entirely on the free databases.

Fixed-Line and Mobile Networks. Figure 4 shows accuracies observed in New York, Chicago, and Philadelphia for major broadband carriers in each market. Again the CDF is the share of location reports. In the best cases, such as either RCN or Comcast on MaxMind in Chicago, the median error is less than 5 km. In each city/database pair, the accuracy is good for fixed broadband and poor for any mobile broadband. This Figure, and others in the main text, rely on ISP classification via whois, as described in Sect. 3.1. A version of this Figure based on an IP addresses' ASNs, is included in the appendix, and is very consistent.

In Chicago, MaxMind is more accurate with fixed-line (AT&T, RCN, WOW, and Comcast) than on mobile (AT&T Mobile, T-Mobile, Sprint, Verizon Mobile) carriers. (IP2Location performs poorly with RCN.) Similarly in New York, Charter, Cablevision, Comcast and Verizon are better localized than AT&T Mobile, Sprint, T-Mobile, and Verizon Mobile; and in Philadelphia, geolocation is more accurate on Comcast than Verizon, T-Mobile, AT&T Mobile, or Verizon Mobile.

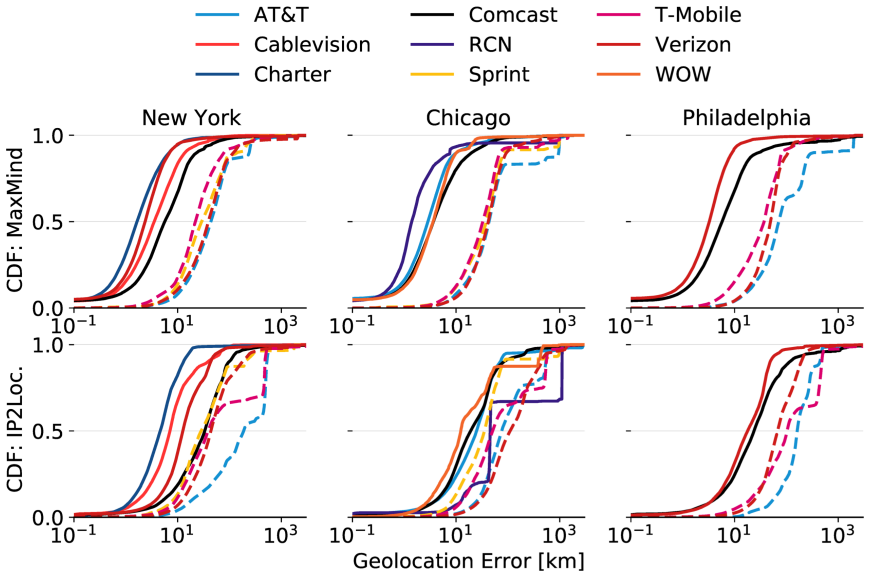

Fig. 4. Geolocation performance by city, database provider, and ISP. Free versions of the database are used in each case. ISPs are shown by their "brand" colors, according to the whois database, which leaves the Sprint and T-Mobile networks distinguishable. Fixed-line networks are denoted by solid lines while mobile networks shown by dashed lines. (Color figure online)

Quantitatively, the share of Comcast data in New York that MaxMind's free service locates within 10 km of the GPS location is 67%. At the other extreme, 87% of T-Mobile location reports from the New York region are IP geolocated to just two distinct locations representing New York itself and Newark; 98% are assigned either to those two, or to one of six other locations in Philadelphia (3), Providence, Boston, and Washington. As a result, only 18% of devices are assigned within 10 km of their true location. In fairness, it must be emphasized that MaxMind does not *claim* to assign these devices within 10 km: almost all of the T-Mobile addresses assigned to the New York and Newark locations are in the 200 km accuracy class.

This basic dichotomy between mobile and fixed broadband is apparent even within ISPs. AT&T offers both services in Chicago, and the CDFs for its fixed-line and mobile services are widely separated. The individual /24 subnets with the largest geolocation errors all belong to the AT&T Mobility organization. In New York and Philadelphia, AT&T only operates mobile networks, and this is reflected in those cumulative distributions. The observation that mobile and fixed-line networks differ may appear obvious once stated, but it need not have been true. Mobile carriers could have constructed networks and CG-NATs with a fixed set of public IP addresses at each antenna. That does not appear to be what they did.

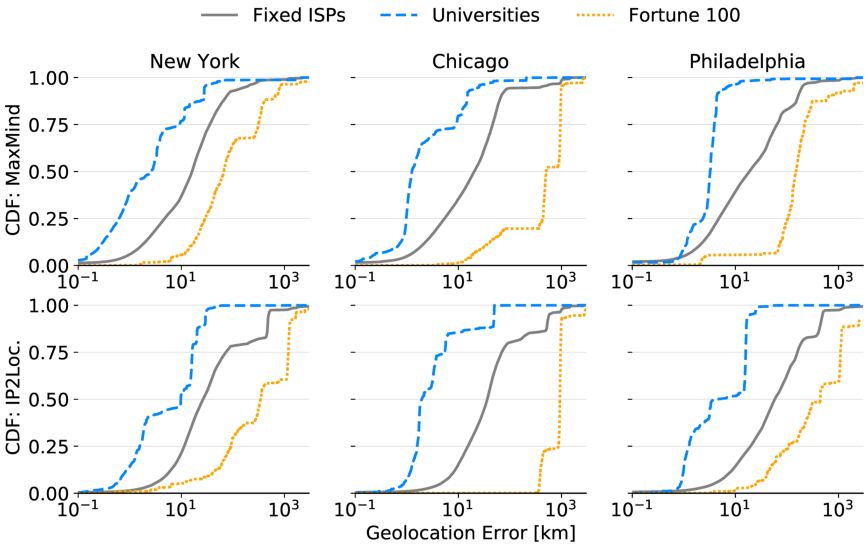

Fig. 5. Geolocation performance on consumer ISPs, contrasted with large universities and Fortune 100 companies.

Universities, Businesses, and Consumer Networks. Before continuing, we also contrast geolocation performance on consumer fixed-broadband, with large universities and companies. We include universities with at least ten thousand students, and Fortune 100 companies other than ISPs. Again, we note that we are implicitly studying the Wi-Fi access points that these institutions operate and which their employees, students, and clients connect to via mobile devices, rather than wired connections or fixed infrastructures of servers. Universities are a classic target in the academic literature on geolocation, but Fig. 5 shows that they are in general more-accurately geolocated than either consumer ISPs or companies. This is not surprising: they have large, physically-concentrated networks, with registration addresses clearly spelled out in ARIN records. In most cases, median geolocation error on MaxMind (free) is less than 2 km, though a few institutes – DePaul in Chicago and the City University of New York – are mislocated by upwards of 10 km. Note that the nominal sample period is August 2020, when students – and indeed many staff and faculty – were not on campus, due to both summer vacation and the coronavirus pandemic.

Figures 3, 4 and 5 suggest that for a substantial share of traffic, IP geolocation is quite accurate. However, this does not do us much good unless those locations can be identified in advance. It is already clear that the picture is rosier with fixed broadband. Those data can be easily identified, either via a `whois` look-up or (in some cases) through the geolocation databases themselves. But mobile and fixed is not the only lever. MaxMind is able to perform better on RCN than on Comcast in Chicago, and better on Charter or Cablevision than Comcast in New York. How are we to identify localizable blocks of addresses?

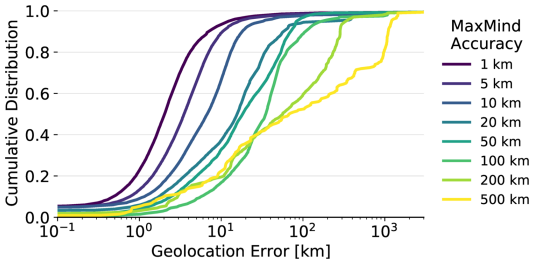

Fig. 6. Cumulative distribution of geolocation accuracy on the MaxMind database, by quoted accuracy bin.

We highlight two additional methods. MaxMind's database provides an "accuracy" field that successfully identifies the precision of entries. Figure 6 shows the CDF for successive bins of claimed accuracy on the free database. In the most precise bin, accuracy of "1 km," the median device in Chicago is geolocated just 2.0 km from the GPS-based location. The "error" with respect to the ground-truth degrades in-line with quoted accuracy, though there is enormous spread in the least-precise, 500 km bin. It is thus *possible* to identify accurately-located addresses – MaxMind does it. But this leaves an open question: *why* are those addresses well or ill-located?

That brings us to the second method. Our hypothesis is that if /24 subnets are geographically localized – small – then addresses within them are more-likely to be accurately geolocated. If they are large, then precise locations would require finer, address-level data. The question can then be re-posed: what is the physical scale of /24 subnets, and is subnet scale in fact correlated with geolocation accuracy?

5.2 What is the Geographic Scale of /24 Subnets?

What are the physical and network properties of accurately-located subnets? In this section, we analyze /24 subnets; in high density cities, where all 255 client addresses *could* credibly be assigned in a small area like a city block. Are they? We require that subnets have at least 10 devices and 10 distinct IP addresses, and focus on a single, fixed network – Comcast. Between the three cities, Comcast has over twenty thousand /24 subnets satisfying these cuts; it carries over 40% of the fixed-line traffic that we observe.

Constructing a Physical Scale. To quantify whether or not a subnet is localized, we define a characteristic physical scale. Many subnets have some outliers, perhaps with locations reported after the fact. To mitigate the impact of these outliers, we must first identify them. We compute the medioid of locations in the subnet, defined in this case simply as the median of the x and y coordinates in a projected (flat) geometry (EPSG 2163). We then measure individual locations' distances from that medioid. We select a configurable fraction f of the data that is "closest" by that measure. For that subset of the data, we calculate the convex

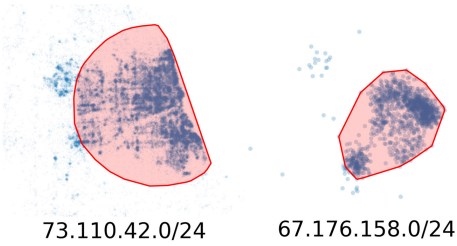

Fig. 7. Illustration of the procedure defining the physical scale of /24 subnets, for one dispersed and one well-localized subnet in Chicago. Convex hulls wrap around $f = 0.9$ of the points within the subnet. The "scale" is the square root of this area. The linear scale on the right-hand side (67.176.158.0/24) is a factor of 8 larger than on the right-hand side. Gaussian noise has been added to the locations for illustrative purposes only.

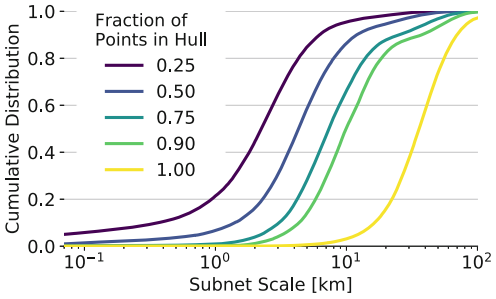

Fig. 8. Cumulative density of subnets' distance scale as derived from the convex hull of locations, as described in the text.

hull. If $f = 1$, then the convex hull covers all locations recorded on the subnet; if $f = 1/2$, it covers the half of points closest to the medioid. Finally we take the area of the convex hull, and "convert" this area to a distance by taking its square root. That square root defines the length scale of the subnet. Figure 7 illustrates this procedure for two subnets. (To preserve anonymity, random noise has been added to the individual points in the illustration.)

Figure 8 shows this distance scale for /24 subnets with at least 10 devices and addresses, for several choices of f. By construction, the scale is smaller or larger when outliers are more or less suppressed, respectively. Setting $f = 0.5$ results in a median subnet scale of 4.3 km, and $f = 0.9$ leads to a scale of 9.9 km. However, the proportion of subnets with scales exceeding 10 km is small for any choice of $f < 0.9$.

The Relationship of Physical Scale and Accuracy. Armed with this scale, we return to the earlier question: when can *addresses* be accurately located? Discarding locations with geolocation error over 100 km, the correlation is 0.69 between the $f = 0.75$ scale of /24 subnets and mean address geolocation error, for

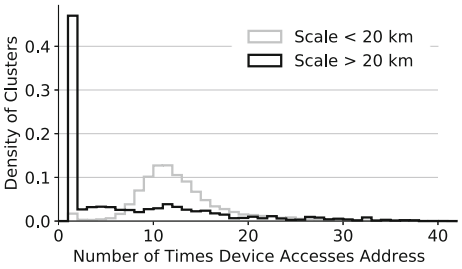

Fig. 9. The number of times a single device visits a single IP address on the subnet (weighted by visits). On subnets with scale greater than 20 km ($f = 0.75$), nearly half of visits device/IP pairs are unique.

MaxMind Free (GeoLite). However, that correlation is only 0.30 on IP2Location, which has worse performance overall. We thus confirm the hypothesis that localization and localizability are related, though strictly speaking, this analysis is not causal.

Still, this analysis has delayed but not *answered* the question; it suggests that geolocation fails on fixed-line addresses when their /24 subnets are geographically dispersed, but this in turn raises the issue of why these disperse subnets exist at all. Comcast has /24 subnets that are spatially concentrated and others that are disperse. Are disperse ones used differently?

We hypothesize that the spatially-concentrated subnets are nearly static whereas large ones provide a reserve of "ephemeral" addresses – perhaps for devices awaiting assignment of a long-term address. A client assigned to an "ephemeral" address would be unlikely to fall on that same address again, whereas a "sticky" address granted to a home network would be used repeatedly. The relevant variable is thus the number of times that a single client is observed at each IP address (weighted by visits). Figure 9 confirms the hypothesis: for subnets with scale greater than 20 km ($f = 0.75$) nearly half of visitors to an IP address visit exactly once.

This behavior is reproduced on Charter, Cablevision, and RCN. It is true to a lesser extent on AT&T, in the sense that devices register far fewer locations on addresses from physically-disperse subnets than on concentrated ones, but the mode at a single visit is not present. Verizon and WOW do not reproduce this behavior.

5.3 How Persistent are the Physical Locations of /24 Subnets?

Geolocation providers are quick to point out that databases evolve continuously. Clearly, the physical infrastructure of the Internet evolves over time, but how quickly do subnets actually move? Because mobile networks subnets are already physically very large, and addresses on them are not accurately located, we focus this analysis on fixed-line broadband.

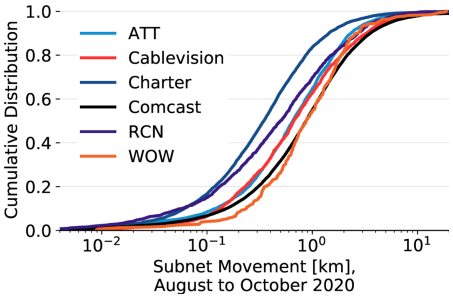

Fig. 10. Distance moved by the medioids of /24 subnet on fixed-line networks, over a two-month period from August to October 2020.

The Movement of Subnets. Figure 10 presents the physical distance between the medioids of individual /24 subnets, as constructed in August and October 2020. As in Sect. 5.2, the medioid is the median of the x and y coordinates. To enter into this figure, subnets must have at least ten unique devices and ten unique addresses in each month. We consider only fixed-line broadband carriers, for this exercise.

On each network considered, the median /24 subnet moves less than a kilometer; There is some inherent variability in our construction of the medioid as the "location" of the subnet in each period, and the Figure shows the difference of these two "noisy" measurements. We thus suspect that this overstates movement. In short, we conclude that on this time scale, subnet locations are quite stable.

Is the Sample Biased? A substantial threat to this analysis is sample composition: by requiring 10 devices and 10 addresses, the subnet *must* be observed in New York, Chicago, or Philadelphia in both months, to enter the sample at all. However, it does not seem to be the case that subnets are moving out of sample. Of the subnets satisfying the cuts in August, 92% also pass them in October (vice versa, 96%). If we raise the thresholds to enter the sample, requiring 20 devices and 20 addresses, 95% of /24 subnets passing these cuts in August also show up with at least 10 devices in October (vice versa, 98%). Raising the thresholds yet further to 50 devices and 50 addresses, the persistence from August to October exceeds 99% (vice versa, 98%).

5.4 How Long Does a Consumer Connection Retain an IP Address?

The analyses above show that IP addresses identify physical locations at the level of 2 km, under the best circumstances. On its own, the IP address clearly does not identify individuals.

Of course, physical locations – geographic coordinates – are not the only way in which IP addresses identify people. Linked to log-ins or other online behaviors, IP addresses can be used to track users over time even without cookies

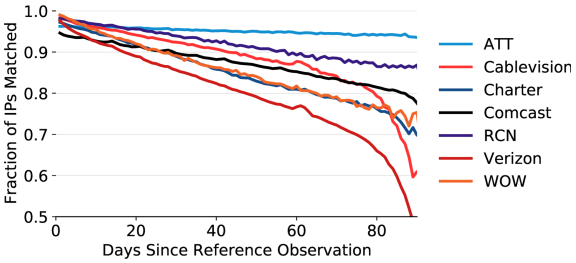

Fig. 11. Persistence of IP addresses. The Figure shows the share of night-time clusters on a single ISP and device, separated by d days, for which the IP addresses are equal on both clusters. Note that for visual clarity, the y axis begins at 0.5 instead of 0.

or fingerprinting (or as a component of a fingerprint). If the IP address is static for a long time, it easier to link online behaviors. A critical concern is thus *how long* fixed-line IP addresses remain with a single household.

Defining Churn. We define *churn* as the likelihood of a device returning to the same IP address on an ISP, after a delay of d days. The denominator includes every pair of night-time connections by a single device to one ISP, d days apart. We select night-time activity, to focus on periods when devices can be reasonably assumed "at home." The numerator is the number of those pairs for which the two nights' connections are on the same IP address. Stated less formally: if I see a device on Monday night ($d = 0$) and again on the same ISP Tuesday night ($d = 1$), what are the chances that it will be on the same IP address? What about next Monday ($d = 7$)?

Since the sample selection is somewhat peculiar – devices are necessarily recorded on fixed-line broadband on multiple nights – one should take some care in interpreting these results. This consideration is particularly acute at the maximum of the range, since there are fewer opportunities for a device to be observed 80 days apart (just 10) let alone 90 (just 1). This perhaps explains the drop-off on the right-hand side.

Rates of Change, Over Two Months. Figure 11 shows the persistence of IP-addresses on fixed-line broadband ISPs. It is clear that devices "leave" individual IP addresses gradually, but at different rates on different ISPs. After one month, more than 90% of devices observed reconnecting to AT&T, RCN, and Cablevision do so on the same IP address. After two months, more than three-quarters of devices return to the same IP address, for all major ISPs in the three cities shown.

6 Can IP Geolocation Databases be Used to Study Internet Access?

At this stage, we would usually turn to a general discussion of findings. Here, we focus our discussion and extend our results, according to the question that

originally motivated our work: assessing the potential for using IP-referenced data in *social science* research on Internet access. *Where* and *for what demographic groups* is geolocation accurate? Can *IP geography* enable *Internet demography*? To make this query concrete, imagine a study of the "homework gap" – (in)equity in access to digital resources for education – based solely on server logs from a site like Wikipedia. If we observe frequencies of use by IP subnet *alone*, can we infer what groups do and do not access the site?

General Considerations. This question is non-trivial, since it confronts the correlations of population density and demographics with geolocation accuracy, along with the spatial patterns of connection modality (mobile vs fixed). Cities have smaller subnets simply because they have higher density of people and devices. They also tend to have larger minority populations. This alone leads to a correlation between geolocation accuracy with demographics or disadvantage. For Chicago and its buffer, the correlation between tract median geolocation error on MaxMind (free) and population density is -0.09 ($p < 0.0001$); in turn, population density is correlated with log median household income ($r = -0.18$, $p < 10^{-10}$). Both of these are small but significant. The flip side of better accuracy at higher density is that distance precision *has* to improve in dense environments, to associate activity with the right population. It's easier to "jump" over many people when they are close together.

Accuracy also varies *within* the city, due to heterogeneity in the fraction of people on mobile vs fixed broadband. There are two reasons for this. People use mobile devices (1) when they are on the go, or (2) because they do not have access to a fixed broadband connection at home. That means that devices in the present sample observed in city centers appear to have "inaccurate" IP geolocation, simply because the device users are more-likely on mobile on the way to or at work. On the other hand, populations without fixed broadband access are unlikely to be accurately IP geolocated, even in their home neighborhood.

As a final consideration before proceeding, one must not confound "unknown" addresses with "mis-located" ones. For example, if a default database location for T-Mobile addresses sits in a particular neighborhood, that neighborhood will appear to have "accurate" geolocation, even though the locations are not known any better than elsewhere. Performance will appear to "degrade" radially, with distance from the default location. Since the default locations are usually in or near cities, that would (*ceteris paribus*) give a false impression that IP addresses in cities (or near the center of the United States, for instance) are accurately-located.

Differences in Access Modality by Demographic Group. Returning to the data, Fig. 12 presents the proportion of the night-time clusters in each tract of Chicago, that are on fixed and mobile broadband. Note that the data are inherently mobile devices with GPS chips; this does not include laptops, for instance. This classifies AT&T, Comcast, WOW, and RCN, as fixed-line providers, and T-Mobile, Sprint, Verizon, and AT&T Mobile as mobile. For those familiar with Chicago, the results are no surprise: the proportion of night-time pings on mobile networks is lower on the wealthier North Side of the city than on the West

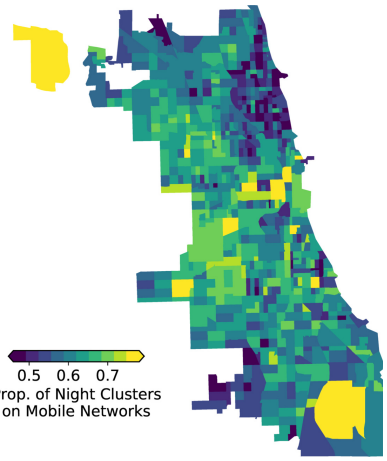

Fig. 12. Proportion of night-time clusters in Chicago recorded on mobile networks.

or South Sides. Indeed, our eyes do not deceive us: the tract level correlation between this constructed variable and share of households with a broadband contract as reported to the Census is -0.25. The correlation between the proportion of night-time pings on mobile networks and the proportion of a neighborhood that is Hispanic is 0.23 (both $p < 10^{-10}$). In other words, connection type is correlated with demographic factors and broadband adoption. This would be reflected in geolocation accuracy. In practice, this means that limiting analyses of Internet activity to accurately-located, fixed-line IP addresses would disproportionately drop traffic from lower-income and minority populations.

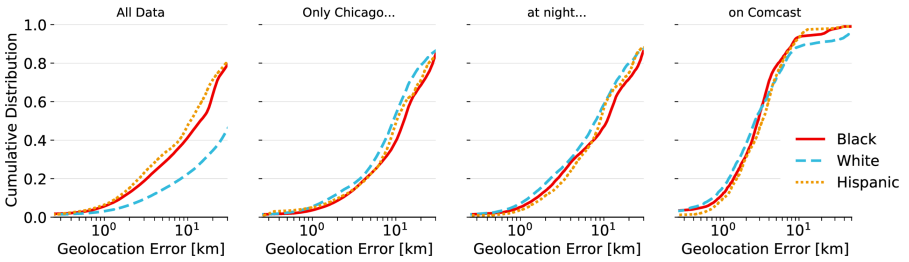

Fig. 13. Cumulative distribution of geolocation error for tracts with white, Black, and Hispanic super-majorities. The first panel presents all data, while the second through fourth restrict to Chicago, Chicago at night, and Chicago at night on Comcast.

The Influences of Density, Demographics, and Modality on IP Geolocation Accuracy. Figure 13 offers an alternative view of this effect, disentangling the countervailing forces of density, demographics, and access modality.

It displays the CDF geolocation accuracy in hyper-segregated neighborhoods of Chicago – ones where two-thirds of residents are white (only), Black (alone or in combination with other races), or Hispanic (of any race). These classifications are made based on data from the US Census' American Community Survey (ACS). Moving from left to right, we begin from the full dataset and layer the cumulative requirements of devices in Chicago proper (not the 40 mile buffer), at night (that is, likely at home), and on Comcast (i.e., on a single, fixed broadband network). The CDF shows the share of location reports. The first plot shows an enormous difference between geolocation in "white" tracts and other segregated tracts – geolocation performs much worse. This effect appears to have more to do with density than race: it reverses when focusing on the City of Chicago, and zeroing in on a single network, the performance lines up quite closely. The exception is at the very high end (above 10 km and 90% of the CDF), where there is apparently an error for locations reported from "white" tracts. About 80% of points are within 5 km of the true location, for all three categories of neighborhood.

Attenuation Bias, from Reliance on Mis-Attributed IP Addresses. The analyses of device modalities above suggest that IP geolocation databases' ability to attribute online behaviors to populations will tend to fail more often for disadvantaged groups. Still, if we were to persist, what errors might we expect to "accrue," by moving an observation from its GPS-based location to the IP-based location? In essence, this question pits the scale of geolocation accuracy against the physical scale of demographic segregation. If IP geolocation moves a point among communities with similar demographics, the error does not directly bias results.

This illustrative analysis is limited to fixed-broadband data from Comcast, where geolocation has a chance of succeeding. Figure 14 presents the log median household income as it would be imputed from a MaxMind look-up, against the true median household income of the neighborhood (Census tract). This results in an unsurprising regression to the mean: as is the usual case with measurement error, the slope is simply attenuated. This suggests that even for fixed broadband, efforts to use IP address alone to "link" online behaviors with human populations are inadvisable at this physical scale. They will in general yield estimates whose magnitudes are biased down. In other words, measurements of "who uses what" that rely on IP geolocation will tend to *understate* differential access. This is consistent with Ganelin and Chuang's work on the socioeconomic status of MOOC registrants. They found that using IP geolocation to identify users' neighborhoods led to underestimates of inequity in adoption [7].

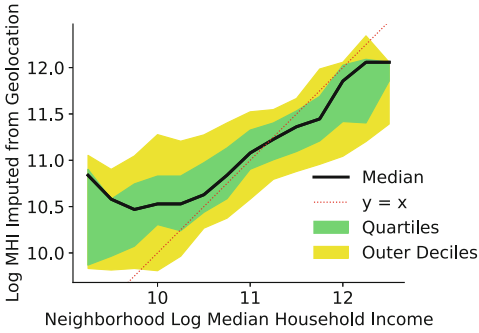

Fig. 14. Quantiles of neighborhood log median household income as "imputed" from MaxMind geolocation (y) as a function of the true neighborhood value (x).

7 Conclusion

Using a large sample of GPS-based smartphone locations this paper has quantified the performance of commercial geolocation databases in New York, Chicago, and Philadelphia. The precision of this analysis far outstrips past work. The analysis has demonstrated significant heterogeneity in geolocation accuracy. The median error for MaxMind's free service is well less than 10 km on fixed commercial broadband networks and at Universities. On mobile networks, IP geolocation is not accurate below the city level. While we consider that consumer devices in large cities in the United States represents a particularly useful vantage point, our conclusions concerning database accuracy and network structure are necessarily limited to the setting that we have observed.

Our analysis has also sought to explain *why* some addresses are accurately located whereas others are not. The physical size of/24 subnets is strongly correlated with accuracy. Geographically disperse/24's appear to be used for "ephemeral" addresses, which clients do not use repeatedly.

Finally, we have contextualized these findings for applications to research on human populations. Both the present data and existing surveys show that disadvantaged populations are less likely to use a fixed broadband subscription at home. Traffic originating from mobile broadband networks cannot be accurately attributed to a neighborhood-level geography, and dropping this traffic altogether would disproportionately remove from analysis the traffic associated with poorer populations. Focussing on the fixed-line context where geolocation is more reliable, the accuracy is still inadequate for associating online activities with real-world geographies and demographics.

From a privacy perspective, a single IP address does not identify an individual, but it both localizes private networks and provides an "index" through time that may be used to aggregate other indirect identifiers. We have shown that the time for IP reassignment of fixed-line broadband consumers varies by ISP, but is typically on the order of months.

A Additional Plots and Tables

See Tables 4, 5 and Figs. 15, 16.

Table 4. Proportion of bumps and clusters according to the classification type assigned by Unacast (cf Sect. 3.1).

Cluster class	Bumps	Clusters
Long Area Dwell	0.382	0.079
Travel	0.322	0.264
Area Dwell	0.193	0.173
Short Area Dwell	0.074	0.290
Potential Area Dwell	0.025	0.127
Ping	0.003	0.066
Large Variance	0.001	0.000
Moving	0.000	0.000
Split	0.000	0.000

Table 5. Proportion of clusters with IP addresses in foreign Internet registries (cf Sect. 3.1).

NIC	Frac.
AFRINIC	0.00011
APNIC	0.00015
LACNIC	0.00016
RIPE	0.00169

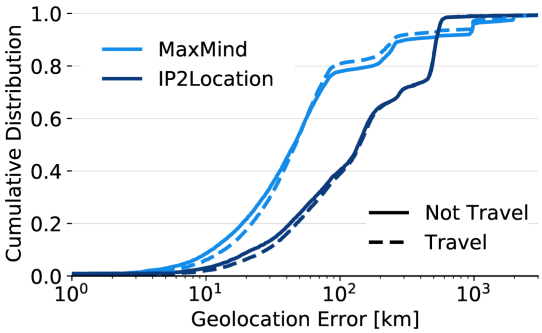

Fig. 15. Empirical cumulative distribution of geolocation accuracy, for travel and non-travel clusters on AT&T's mobile network, as evaluated on the free versions of the MaxMind and IP2Location databases (cf Sect. 4.1).

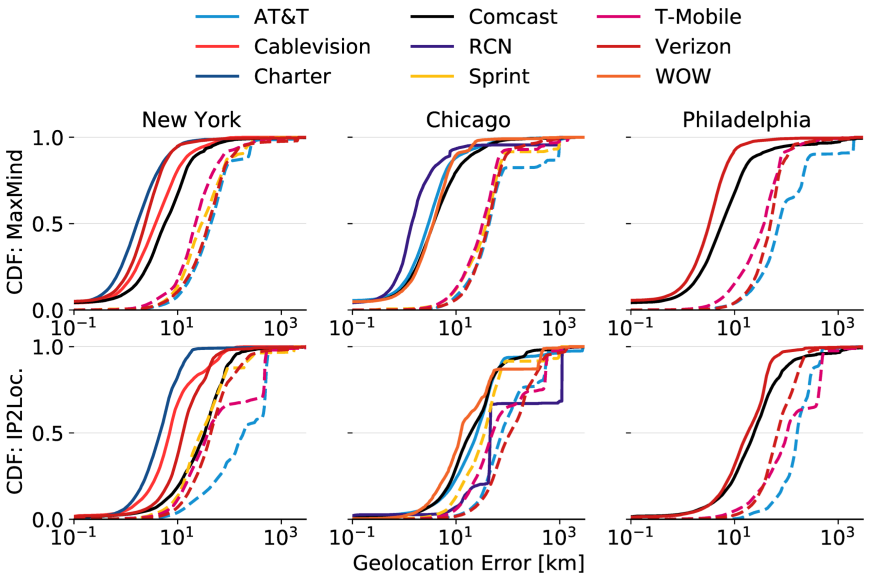

Fig. 16. Geolocation performance by city, database provider, and ISP. The Figure is identical to Fig. 4 of the text, except that ISPs are identified by ASN instead of via whois. ASNs associated with each ISP are listed in Table 6. Free versions of the database are used in each case. ISPs are shown by their "brand" colors, according to the whois database, which leaves the Sprint and T-Mobile networks distinguishable. Fixed-line networks are denoted by solid lines while mobile networks shown by dashed lines. (Color figure online)

Table 6. Autonomous systems asssociated with each ISP, for the data within the study region. This listing is a categorization of the ASNs seen most-frequently in the data. It is not expected to be an exhaustive listing of all ASes corresponding to the ISPs, even in the New York, Chicago, and Philadelphia regions. ASNs are ordered by the number of /24 subnets observed in the data.

ISP	ASNs
AT&T	7018, 2386, 6389, 2686, 4473, 4466, 797, 6431, 17225, 17227
AT&T Mobile	20057
Cablevision	6128, 13490, 32953, 14638, 19720
Charter	12271, 10796, 20115, 11351, 11426, 33363, 20001, 11427, 33588, 14065, 7843, 17359, 16787
Comcast	7922, 33491, 33659, 33287, 7016, 33657, 33651, 7725, 7015, 20214, 33661, 395980, 33652, 396019, 396021
RCN	6079
Sprint	10507, 1239
T-Mobile	21928
Verizon	701, 2828, 23148, 15133, 11486, 12079
Verizon Mobile	22394, 6256, 6167
WOW!	12083, 11693

References

1. Arif, M.J., Karunasekera, S., Kulkarni, S., Gunatilaka, A., Ristic, B.: Internet host geolocation using maximum likelihood estimation technique. In: 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 422–429. IEEE, Perth (2010)
2. Dan, O., Parikh, V., Davison, B.D.: IP geolocation through reverse DNS. CoRR abs/1811.04288, pp. 1–10 (2018)
3. Eriksson, B., Barford, P., Maggs, B., Nowak, R.: Posit: a lightweight approach for IP geolocation. SIGMETRICS Perform. Eval. Rev. **40**(2), 2–11 (2012). https://doi.org/10.1145/2381056.2381058
4. Eriksson, B., Barford, P., Sommers, J., Nowak, R.: A learning-based approach for IP geolocation. In: Krishnamurthy, A., Plattner, B. (eds.) PAM 2010. LNCS, vol. 6032, pp. 171–180. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12334-4_18
5. Feldman, D., Shavitt, Y., Zilberman, N.: A structural approach for pop geo-location. Comput. Netw. **56**(3), 1029–1040 (2012). https://doi.org/10.1016/j.comnet.2011.10.029. http://www.sciencedirect.com/science/article/pii/S1389128611004191
6. Freedman, M.J., Vutukuru, M., Feamster, N., Balakrishnan, H.: Geographic locality of IP prefixes. In: 5th ACM SIGCOMM Conference on Internet Measurement, IMC '05, pp. 153–158. USENIX Association, USA (2005). https://doi.org/10.5555/1251086.1251099
7. Ganelin, D., Chuang, I.: IP geolocation underestimates regressive economic patterns in MOOC usage. In: 11th International Conference on Education Technology and Computers, pp. 268–272. Association for Computing Machinery, New York City (2019). https://doi.org/10.1145/3369255.3369301
8. Gharaibeh, M., Shah, A., Huffaker, B., Zhang, H., Ensafi, R., Papadopoulos, C.: A look at router geolocation in public and commercial databases. In: Internet Measurement Conference, IMC '17, pp. 463–469. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3131365.3131380
9. Gill, P., Ganjali, Y., Wong, B., Lie, D.: Dude, where's that IP? circumventing measurement-based IP geolocation. In: 19th USENIX Conference on Security, USENIX Security'10, p. 16. USENIX Association, USA (2010). https://doi.org/10.5555/1929820.1929842
10. Gouel, M., Vermeulen, K., Beverly, R., Fourmaux, O., Friedman, T.: IP geolocation database stability and implications for network research. In: Proceedings of the Network Traffic Measurement and Analysis (TMA) Conference. Online (2021). https://www.cmand.org/papers/geostable-tma21.pdf
11. Gueye, B., Uhlig, S., Fdida, S.: Investigating the imprecision of IP block-based geolocation. In: Uhlig, S., Papagiannaki, K., Bonaventure, O. (eds.) PAM 2007. LNCS, vol. 4427, pp. 237–240. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71617-4_26
12. Gueye, B., Ziviani, A., Crovella, M., Fdida, S.: Constraint-based geolocation of internet hosts. IEEE/ACM Trans. Netw. **14**(6), 1219–1232 (2006). https://doi.org/10.1109/TNET.2006.886332
13. Hu, Z., Heidemann, J., Pradkin, Y.: Towards geolocation of millions of IP addresses. In: Internet Measurement Conference, IMC '12, pp. 123–130. Association for Computing Machinery, New York (2012). https://doi.org/10.1145/2398776.2398790

14. Huffaker, B., Fomenkov, M.: kc claffy: geocompare: a comparison of public and commercial geolocation databases. Technical report, Cooperative Association for Internet Data Analysis (CAIDA), San Diego, CA (2011). https://www.caida.org/publications/papers/2011/geocompare-tr/
15. Katz-Bassett, E., John, J.P., Krishnamurthy, A., Wetherall, D., Anderson, T., Chawathe, Y.: Towards IP geolocation using delay and topology measurements. In: 6th ACM SIGCOMM Conference on Internet Measurement, IMC '06, pp. 71–84. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1177080.1177090
16. Komosny, D., Rehman, S.U.: Survival analysis and prediction model of IP address assignment duration. IEEE Access **8**, 162507–162515 (2020). https://doi.org/10.1109/ACCESS.2020.3021760
17. Lab, M.: M-lab visualizations (2021). https://www.measurementlab.net/visualizations/
18. Lazer, D.: Computational social science. Science **323**(5915), 721–723 (2009). https://doi.org/10.1126/science.1167742
19. Li, D., et al.: IP-geolocation mapping for moderately connected internet regions. IEEE Trans. Parallel Distrib. Syst. **24**(2), 381–391 (2013). https://doi.org/10.1109/TPDS.2012.136
20. Mishra, V., Laperdrix, P., Vastel, A., Rudametkin, W., Rouvoy, R., Lopatka, M.: Don't count me out: On the relevance of IP address in the tracking ecosystem. In: Proceedings of The Web Conference 2020, WWW '20, pp. 808–815. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3366423.3380161
21. Padmanabhan, R., Dhamdhere, A., Aben, E., Claffy, k., Spring, N.: Reasons dynamic addresses change. In: Proceedings of the 2016 Internet Measurement Conference, IMC '16, pp. 183–198. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2987443.2987461
22. Padmanabhan, R., Rula, J.P., Richter, P., Strowes, S.D., Dainotti, A.: DynamIPs: analyzing address assignment practices in ipv4 and ipv6. In: Proceedings of the 16th International Conference on Emerging Networking EXperiments and Technologies, CoNEXT '20, Association pp. 55–70. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3386367.3431314
23. Padmanabhan, V.N., Subramanian, L.: An investigation of geographic mapping techniques for internet hosts. In: Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. SIGCOMM '01, pp. 173–185. Association for Computing Machinery, New York (2001). https://doi.org/10.1145/383059.383073
24. Pereira, M., Kim, A., Allen, J., White, K., Ferres, J.L., Dodhia, R.: U.S. broadband coverage data set: a differentially private data release, pp. 1–7 (2021)
25. Poese, I., Uhlig, S., Kaafar, M.A., Donnet, B., Gueye, B.: IP geolocation databases: unreliable? SIGCOMM Comput. Commun. Rev. **41**(2), 53–56 (2011). https://doi.org/10.1145/1971162.1971171
26. Richter, P., et al.: A multi-perspective analysis of carrier-grade NAT deployment. In: Proceedings of the 2016 Internet Measurement Conference, IMC '16, pp. 215–229. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2987443.2987474
27. Shavitt, Y., Zilberman, N.: A geolocation databases study. IEEE J. Sel. Areas Commun. **29**(10), 2044–2056 (2011). https://doi.org/10.1109/JSAC.2011.111214

28. Spring, N., Mahajan, R., Wetherall, D.: Measuring ISP topologies with Rocketfuel. In: Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM '02, pp. 133–145. Association for Computing Machinery, New York (2002). https://doi.org/10.1145/633025.633039
29. Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., Huang, C.: Towards street-level client-independent IP geolocation. In: 8th USENIX Conference on Networked Systems Design and Implementation, NSDI'11, USA, pp. 365–379 (2011). https://doi.org/10.5555/1972457.1972494
30. Wong, B., Stoyanov, I.: Octant: a comprehensive framework for the geolocalization of internet hosts. In: 4th USENIX Symposium on Networked Systems Design & Implementation (NSDI 07), pp. 313–326. USENIX Association, Cambridge (2007). https://www.usenix.org/conference/nsdi-07/octant-comprehensive-framework-geolocalization-internet-hosts
31. Youn, I., Mark, B.L., Richards, D.: Statistical geolocation of internet hosts. In: 18th International Conference on Computer Communications and Networks, pp. 1–6. IEEE, San Francisco (2009). https://doi.org/10.1109/ICCCN.2009.5235373