Geo-Network Operations at Cloudflare

Joe Abley, Cloudflare, jabley@cloudflare.com

Cloudflare supports the development of an open, interoperable, secure and complete framework for the management of geolocation data and stands ready to contribute and participate.

"Where is this IP address?"

Even if we assume that we have a consistent, shared understanding of what “where” means (lattitude/longitude? postal code? city? region?), the innocent-looking phrase "where is this IP address?" is actually a surprisingly complex question.

The most useful answer usually depends on the context for the question. For example, are we asking because we want to optimise some delivery of content for highest performance? Or are we really asking "am I permitted to serve content to this user?" or even "what regionalisation settings are appropriate for this request"? The same question asked for different reasons might lead to different answers. And then, even when the question is unambiguous and even when we know what kind of answer is needed, finding answers is complicated.

The topology of the Internet has no centralised control. This is a crucial and necessary ingredient of its success, but a consequence is that the topology is never static. Multiple such dynamic topologies can be found layered upon each other, autonomously-operated topology stacks are deployed at will and managed independently and simultaneously and endpoint identifiers can move with the speed at which each of those layers can be reconfigured, both under provider and end-user control. Networks that forward IP datagrams require interface addresses attached to their topologies, so if topologies can change frequently we expect addresses to move too. The existence of an active resource leasing and transfer market for IPv4 addresses adds another dimension of mobility. Any dataset that maps address to geography is at best a snapshot in time and a fuzzy representation of a small handful of vantage points.

Most questions of address locality are dispatched to a single source of answers, in the form of a single tool, a single dataset or a single data provider. Multiple of all of these exist, but their identities are often opaque. In particular, in our experience a single, long-established and dominant data provider is often assumed to be the single source of authority, despite the existence of others.

The impact of getting the answer wrong varies and can be serious. A web page served from a more distant source might take slightly longer to load, or flash crowds of mis-located clients might cause congestion and network failure; landing pages might be presented in languages that are unfamiliar to users; content might be unavailable for licensing reasons; connections

might be refused when they ought properly to be allowed. The impact is generally distributed and difficult to measure across a meaningful sample population.

Address Geolocation at Cloudflare

Cloudflare provides many services over the Internet including CDN services, network access services, privacy-centric VPN services and a distributed edge compute platform known as Cloudflare Workers. Almost all network services at Cloudflare make extensive use of anycast.

Outbound connections from Cloudflare's network are associated with many different services, including origin fetches in the CDN, connections originated by VPN clients and connections originated by on-premises hosts who use Cloudflare for connectivity. Cloudflare maintains a geolocation dataset for the addresses used to source outbound connections internally, populated with customer data and published[1] in the format specified in RFC 8805.[2] This dataset is consumed by numerous IP geolocation data brokers.

Many Cloudflare services can be configured to use IP address geolocation to configure how inbound connections from particular addresses should be handled, including (broadly) CDN services and customer workloads deployed in Cloudflare Workers. Cloudflare consumes geolocation datasets under licence from an external geolocation data broker to provide the basis for such operational decisions.

Gaps in Existing Guidance and Related Problems

Authentication. RFC 9632[3] describes an optional mechanism for authenticating geofeed data using RPKI validation. We are not aware of a meaningful example of this mechanism being used; in our experience most network resources that are able to be validated using the RPKI feature signatures from keys that are not available to the resource holders themselves, but are rather maintained by RPKI facilitators such as RIRs. These keys are therefore not generally available to sign arbitrary objects. Aside from these practical considerations and the acknowledged limitations in the ability to authenticate RPSL objects, individual rows in a published dataset might well relate to different authorised parties, and there is no granular mechanism available to validate individual mapping assertions. We think RFC 9632 is a good start to a conversation, but that more is needed.

Integrity Protection. There is no reliable integrity protection of mappings published by the operator of a network, and no ability to understand the provenance of any particular mapping. A consumer of a set of mappings receives the results of an unknown number of corrections and transformations made by unknown parties for unknown reasons.

---

[1] https://api.cloudflare.com/local-ip-ranges.csv
[2] Kline, E., Duleba, K., Szamonek, Z., Moser, S., and W. Kumari, "A Format for Self-Published IP Geolocation Feeds", RFC 8805, DOI 10.17487/RFC8805, August 2020.
[3] Bush, R., Candela, M., Kumari, W., and R. Housley, "Finding and Using Geofeed Data", RFC 9632, DOI 10.17487/RFC9632, August 2024.

Provenance. It is common for assertions about the locations of particular addresses to be "corrected" between original publisher and their ultimate consumer, e.g. by means of active measurements or by inferences drawn from other published location information such as data published by RIRs. Since there is no ability to determine the provenance of any particular mapping, it is usually not possible for relying parties to identify or to reverse individual corrections in an informed way.

Consistency. Different data brokers often publish different locations for the same address, e.g. because their internal processes for corrections and for prioritising particular inferences over others are different.

Data Quality. Data quality is a recurring problem. All of the location-related metadata for addresses specified in RFC 8805 is encoded as unstructured, free text. The problem that there are often multiple, correct ways to represent the name of a place is unaddressed. The interpretation of metadata that is ambiguous, or mistyped, or of files that are malformed in some other way is not specified, and there is no canonical guidance for validation of a dataset before publication to ensure that its contents are likely to be consumable. We appreciate that this is a hard problem, but we think it's also a problem that might well have found solutions elsewhere[4] and we think these things are worth thinking about.

Supportability. Cloudflare's customer support is frequently asked to fix particular published geolocation mappings. Customers whose services are attached to addresses assigned by Cloudflare have a reasonable expectation that we can and will fix these problems; in practice, the gaps described above mean that the extent to which Cloudflare can help is limited and the impact of any action taken by Cloudflare is difficult to predict. The effects of incorrect mappings can persist despite accurate data being published by Cloudflare, and the path to a solution is often either unclear or revealed only through guesswork.

Future Opportunities

The various possible geographic locations that are attributed to a single address are examples of metadata that could be published by the address's authenticated, legitimate operator, providing clear and compelling signals to devices elsewhere about the disposition of the address and the expectations around traffic associated with that address. Such signals could be combined with existing, more error-prone signals to improve service quality and reduce the impact of false positives. Tags such as "residential user", "CDN", "VPN endpoint" and "anycast" that are sometimes provided by existing geolocation brokers provide useful illustrations but are themselves usually based on third-party inferences and are known to be error-prone. An extensible means of communicating intent by an authenticated authority would have application in many areas and has the potential to dramatically improve the stability and security of Internet traffic.

[4] For example, a related problem in trade and transport led to the development and maintenance of UN/LOCODE, a set of over 110,000 unique identifiers for functions such as seaports, rail and road terminals, airports, Post Exchange Offices and border crossing points.