Country-in-the-Middle: Measuring Paths between People and their Governments

Paper #574, 13 pages body, 17 pages total

ABSTRACT

Understanding where Internet services are hosted, and how users reach them, has captured the interest of government regulators and others concerned with the privacy of data flows. In this paper we focus on government websites—services which arguably merit a higher expectation of protection against foreign surveillance or interference—and seek to identify countries in the middle (CitMs): countries that are neither the source nor destination in a path for a resident visiting their online government services. Finding these CitMs raises daunting methodological challenges. We propose a framework to identify CitMs and use a pilot study of 149 countries to refine our methodology before conducting an in-depth measurement study of 11 countries. For our focused study, we compile an extensive set of websites hosting government services and analyze over 9,000 IP-level paths from vantage points in those countries to these services. We conduct extensive manual validation to corroborate or discard paths based on the aforementioned challenges, and discuss paths that experience unexpected CitMs.

1 INTRODUCTION

Data sovereignty has become increasingly important as countries fear their network traffic may be surveilled—or even tampered with—by foreign nations. Internet traffic that enters a foreign country becomes subject to that country's laws, which may stipulate that the government can gain access under certain conditions. For example, Belarus requires domestic telecom providers to allow various levels of governmental access, including the ability to surveil (potentially foreign) traffic that transits their routers [25]. Such concerns are not hypothetical; in 2022, researchers found that a Russian state-controlled telecom company was routing traffic from Ukrainian networks through Russia [57], and, in 2013, Snowden revealed that GCHQ, a UK intelligence agency, was intercepting traffic on more than 200 fiber-optic cables [36].

Traffic may enter a foreign nation for a variety of political and technical reasons, many of which are benign—for example, due to submarine cables that were built between colonial powers and their former colonies [63]; geographic proximity to other countries; and the use of cloud providers, large transit providers, and Internet exchange points (IXPs). A particular foreign nation that receives traffic may be perceived as positive or negative depending on a country's political relationships. In this work we do not take a stance on the desirability (or lack thereof) of any nations. Regardless of the reason why traffic transits a foreign nation, if a government wants to manage the sovereignty of their data, they must first understand where their traffic is going (even if they ultimately do not change the locality of this traffic). In particular, we focus on one class of Internet traffic where we expect that governments have greater visibility and would like to assess any potential security concerns: traffic between a given country's residents and governmental websites.

Each nation must determine which governmental services to host domestically and which to host abroad. Presumably, such decisions are taken with full knowledge of the identity and applicable regulations of the nation selected to host the service (although the increasing prevalence of IP anycast raises doubts on even that point). It is less clear, however, whether a nation's decision makers are aware of how traffic will be routed between their residents and selected hosting locations. If a government website is hosted within the country itself (which we term a *convergent* destination), it is counterintuitive that a resident's traffic to that website would leave the country. Moreover, even if a website is hosted in a foreign country (which we call a *divergent* destination), it is frequently unclear—and even time-varying—which other countries a resident's traffic may transit.

We consider how to answer the question: when residents located in a given country access their government's websites, what other countries are involved, with what frequency, and why? We refer to any country that is neither the source nor destination country of an IP-level path as a *country in the middle (CitM)*. CitMs can appear in paths to both convergent and divergent destinations. (Prior work sometimes refers to paths to convergent destinations with CitMs as "boomerang" or "tromboning" [13, 14, 18, 43].) Figure 1 depicts how we taxonomize the paths we study.

Despite the alluring simplicity of our question, our repeated attempts over multiple years to answer it have identified five core challenges must be addressed in order to systematically identify CitMs. The first two challenges are compiling a comprehensive set of government websites and finding sufficient in-country vantage points. These challenges are manageable for studies of individual countries (e.g., one could deploy new vantage points, or select only countries with authoritative sources for government websites), but become increasingly daunting as the number of countries

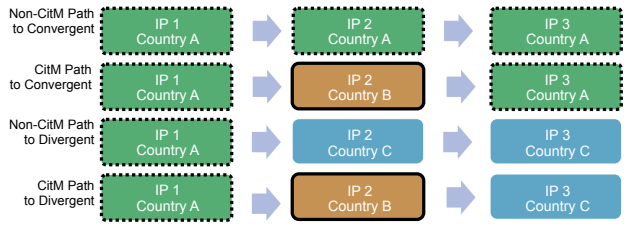


Figure 1: A visualization of our taxonomy. Colors and border patterns indicate the country hosting the client, website, or intermediate hop(s).

increases. Next, any study that relies upon in-band measurement of IP-level paths (which is the only practical approach at scale) must address the reality that not all IP hops are responsive to traceroute or similar tools, raising nuanced questions regarding how to interpret incomplete path information. Finally, identifying CitMs presents two additional challenges: accurately geolocating core infrastructure and identifying anycast destinations, for which it may be difficult to determine the replica in use.

In this paper, we describe how these challenges frustrated our attempts to conduct a global-scale study that collects traceroutes from RIPE Atlas probes located in almost 150 countries around the globe. We use concrete examples from our experience to motivate the design of a rigorous validation methodology that allows us to draw conclusions regarding CitM prevalence in a more focused study. Concretely, we develop a framework to study CitMs and conduct a study of 11 countries in which we use 10 RIPE Atlas probes located in each country to launch traceroutes towards approximately 100 government websites for each country, resulting in a dataset of 9,278 successful[1] traceroutes. We geolocate the intermediate hops to the best of our ability to identify CitMs. Because of the limitations of available geolocation approaches, we cross-validate the geolocation of each alleged CitM hop. While we believe our conclusions are sound, navigating the challenges requires extensive manual effort.

Many of our validated CitMs are plausible yet unexpected, either due to the geography of the countries involved, or because our dataset contains evidence of alternative paths between the same vantage point and the website without CitMs. Operators may have good reasons for routing traffic through other ASes or IXPs, such as to avoid congestion—even if this may introduce CitMs—but we expect that some governments may not be aware of the set of nations through which their traffic may be routed.

The contributions of our work include:

- Detailing the myriad challenges facing a systematic study of governmental traffic sovereignty, developed through a pilot study of CitMs for government websites of almost 150 countries;
- Proposing a framework for studying CitMs;
- Compiling lists of government websites for 11 countries and measuring IP-level paths to these services from in-country vantage points, resulting in a dataset of over 9,000 traceroutes;
- Performing extensive manual validation during data collection and post-processing;
- Extracting insights about common CitMs for the 11 studied countries; and
- Recommending approaches to improve the coverage and reliability of future studies

2 CHALLENGES

In this section, we enumerate the five significant challenges to conducting a study that seeks to uncover threats to traffic sovereignty.

2.1 Compiling government websites

There is no single authoritative source that provides a current, comprehensive list of official government websites for most countries. While the United Nations publishes a list [39], it contains between one and four URLs per country, far too few to constitute a representative sample. Some individual countries publish official lists, but we find many to be outdated and/or inaccurate. In particular, websites on these lists may not actually host content, or—if the list is outdated—may no longer belong to the government.

2.2 Finding vantage points

To measure IP-layer paths, we need vantage points distributed throughout the countries of interest that can run traceroutes to target websites. RIPE Atlas provides many vantage points in a large set of countries (though some countries have few probes) at low cost, but some of these vantage points are deployed in data centers and thus may not be representative of the Internet providers used by residents of a given country. Recent analysis has also found that some RIPE Atlas probes may misreport their geolocation [28], forcing us to scrutinize their purported locations.

2.3 Interpreting traceroute responses

The inferential challenges in interpreting traceroute data are well documented [33]. Routers may filter ICMP traffic, not respond (resulting in unlabeled hops in traceroute outputs), or may respond with bogon (i.e., reserved or unallocated) IP addresses. A more subtle problem is that a router may respond using an IP address of an interface different from that

[1] We fall short of the desired 11,000 traceroutes because some Atlas probes malfunctioned and some collected traceroutes our validation process determines to be unreliable.

which received the packet, or even an IP address that belongs to the interconnecting router for that interface. However, a 2014 study [32] found such responses to be relatively rare, and that traceroute remains the best (only) available tool for path analysis.

Similarly, tunneling techniques like Multiprotocol Label Switching (MPLS) may complicate interpreting traceroute responses by hiding hops inside a tunnel, which in turn may obscure potential CitMs. A 2012 study [9] found that such opaque tunnels that hide hops were rare in the wide-area Internet, and that it was far more common for MPLS tunnels to reveal the routers traversed within the tunnel. That said, cloud providers use such techniques extensively [7], so future work could assess the extent to which our methodology under-reports CitMs due to tunneling in private backbones (Section 9).

2.4 Performing IP geolocation

A fundamental task when attempting to map network paths to countries is geolocating individual IP addresses. Publicly available commercial geolocation databases and research-oriented geolocation services [10, 26, 34, 37, 42] vary tremendously in accuracy, precision, stability, and coverage [8, 10, 16, 17], and are generally not focused on core Internet routers. As a result, these databases must be treated with skepticism and corroborated using other approaches.

2.5 Dealing with anycast

Anycast hosting allows multiple physical servers to present the same public IP address, which poses a challenge to geolocation services that map IP addresses to a single geographic area. Because anycast can obscure the identity of the country hosting the website, it is difficult to determine if a government website is convergent or divergent—or both. Moreover, in the case of divergent destinations, it complicates the validation and interpretation of CitMs because the geographical relationship between the source and destination countries may be unclear. In our work we use an anycast-specific geolocation dataset [20] to identify potential countries hosting the anycast IPs we uncover, but the dataset is unfortunately not contemporaneous with our measurements (see Section 7.2 for details), so we do not incorporate the results into our main findings.

3 BASELINE METHODOLOGY

To begin, we present a(n admittedly incomplete) methodology that navigates as many of these challenges as possible in an automated fashion before or during data collection. We employ this methodology to conduct a global-scale pilot study and use the results to design an enhanced methodology that incorporates manual validation steps where necessary.

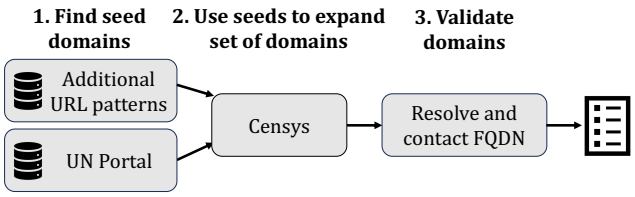


Figure 2: Overview of government domain collection.

3.1 Collecting Government Websites

Given the lack of a single authoritative source, we compile lists of government websites on a per-country basis using a three-step process depicted in Figure 2: (1) We start by combining the few government websites provided by the United Nations [40] with additional sites we obtain from lists of government websites for each country. From each of these URLs, we extract the relevant domain, i.e., eTLD+1, to use as a seed. (2) We identify candidate secure websites within these official domains by consulting the Censys TLS certificate repository [11] and extracting URLs contained within the certificates for seed domains. (3) We arrive at a final set of target domains by confirming the liveness of each candidate website.

3.1.1 Official Government Websites. To expand upon the few sites listed in the UN website, we compile lists of official government websites for each country collected from various sources. The U.S. government publishes expansive datasets with this information [1, 46]. Because manually collecting similar lists for all nations is prohibitively labor intensive, we instead devise an automated approach. After manually correcting some inaccuracies in the U.N. list and adding some well-known domains for other nations we use them to bootstrap a search for additional governmental websites.

3.1.2 Expanding our Set of Domains. Prior work has relied on DNS logs [23] or fixed sets of localized hostname patterns like `*.gov.ccTLD` (and other variations in different languages) to find government websites [31, 60]. Instead, we employ Censys [11] to search for subdomains of our seed domains (and their eTLD+1's) in active TLS certificates. Our approach of extracting eTLD+1's from authoritative domains allows us to not only infer the patterns used in prior work, but also discover eTLD+1's that do not match those patterns. For example, we find Canada's primary government eTLD+1 (`gc.ca`) does not match the patterns used in prior work. Finally, the presence of a TLS certificate suggests that the website operator is following best security practices and that they would like to avoid security concerns like machine-in-the-middle attacks.

3.1.3 Verifying Liveness. The existence of a certificate, however, does not indicate that there is a corresponding operational web service. We confirm the existence of a website by resolving the FQDN and attempting to contact a Web server at that address.

3.2 Measurement Vantage Points

We conduct our measurements from RIPE Atlas probes, a large, distributed set of vantage points for measurement research. We select up to 10 RIPE Atlas probes per country that maximize AS diversity; if a country has fewer than 10 RIPE Atlas probes, we use them all. We will make our code to choose such probes publicly available. Unfortunately, many countries do not have a robust infrastructure of RIPE Atlas probes, reducing our flexibility. We also include data center and anchor probes (which are typically hosted by companies as opposed to individuals); future work should consider whether hardware probes more accurately reflect residential Internet service provider paths.

3.3 Path Collection

We run ICMP (Paris) traceroutes from the selected probes in each country to the set of target domains for that country. (We discuss the implications of using ICMP vs TCP in Section 6.4.2.) We perform DNS resolution on the probe itself to mimic user traffic; this frequently results in different vantage points probing different IP addresses for the same domain. We use IPinfo [26], a geolocation database with higher accuracy compared to others [8], to geolocate each hop including the destination.

4 PRELIMINARY STUDY

We conducted a pilot study in May 2023 that attempted to be as expansive as possible.

4.1 Methodological details

We followed the baseline methodology from the previous section with a few refinements (that we subsequently revise in our focused study as detailed in Section 5.2).

4.1.1 Website filtering. Starting from the United Nations list of websites and a few other seed domains, our Censys search yielded 763,543 FQDNs across 110,068 registered domains, 388 eTLDs, and 196 countries. We filtered the list by using ZDNS [29] with the 8.8.8.8 name server to focus only on domains that successfully resolve. We then used ZMap [12] to ensure port 443 was open on at least one IP address per FQDN, increasing our confidence that there is an operating Web service—but we did not inspect the content available. This filtering process resulted in 497,475 FQDNs spanning 93,548 registered domains and 362 eTLDs.

Resource constraints necessitated sampling among the available FQDNs to select our measurement targets; we sought to maximize the diversity of destination IP addresses. We used the CAIDA prefix2as dataset [5] to find the corresponding IP prefix and AS announcing the IP address associated with each FQDN in our list. We chose up to 100 FQDNs per country that maximize, in order, 1) number of ASes, 2) IP prefix diversity, and 3) number of unique IP addresses. This yielded 12,450 FQDNs for 194 countries. While we attempted to collect 1,000 measurements per country (10 probes each visiting 100 domains) we fell (far) short: our measurement campaign yielded 71,164 traceroutes for 149 countries.

4.1.2 Post-processing. We employ a variety of filtering techniques to increase our confidence in the geolocation data we consider. Within a traceroute, we first strip out all hops in private IP spaces, as it is impossible to get any geolocation information for these addresses (6,633 hops across 64.12% of traceroutes). To reduce noise, we filter hops where the IP address geolocates to a country code that only appears once in the entire traceroute, corresponding to 2,779 hops across 17.50% of traceroutes. We then remove any traceroutes where, as a result of the previous filters, we are left with no hops between the source and destination IP addresses; this is the case for 1,042 traceroutes. After this filtering process, we are left with 68,764 traceroutes for 149 countries.

4.2 Initial results

Figure 3 presents the results of our pilot study, where each country is represented by a stacked horizontal bar chart that plots paths to convergent targets to the left of zero on the x axis, while divergent targets are plotted to the right. Traceroutes to anycast targets (identified using the MAnycast^2 dataset [61] collected on January 2022) are graphed separately on the far right. The bars are color-coded by the number of CitMs in the path and normalized according to the number of probes used in that country, so the length of the bar corresponds to the average number of traceroutes collected per probe—which is expected to be 100, but is frequently less due to measurement errors and/or lack of target domains for that country. Yet, one immediate takeaway from the pilot study is that even for countries where our Censys-based target selection is unable to generate many targets (i.e., those countries with very short bars), we still uncover CitMs (e.g., Liechtenstein, Vanuatu, Montenegro, etc.).

At a high level, these results are consistent with intuition: for example, in general there are more CitMs on paths to divergent targets than convergent ones (i.e., the bars to the right of zero have more pink and orange than those to the left). That said, we do indeed find CitMs on paths to convergent targets, suggesting that even if a government hosts their

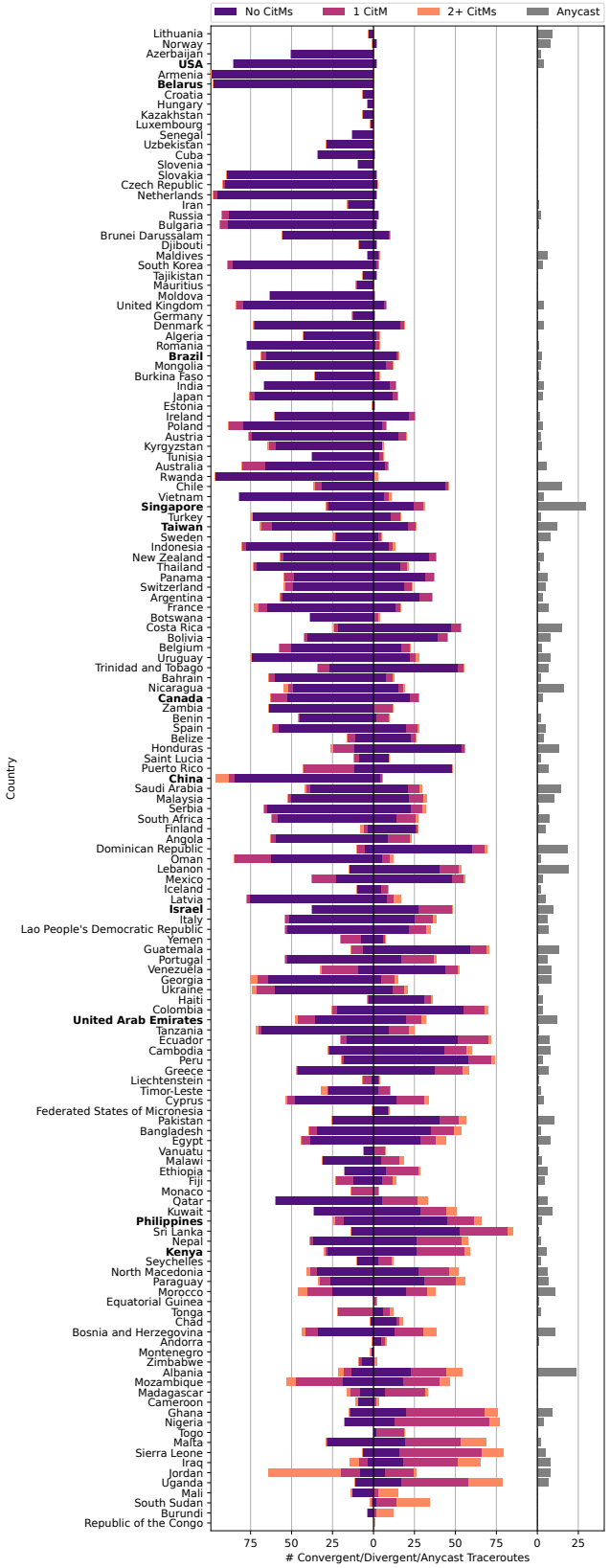


Figure 3 (left): Frequency of CitMs for each country in our pilot study. Traceroutes are normalized by the number of vantage points. Bolded countries indicate countries of focus in our subsequent study.

online services domestically, there is no guarantee that traffic from residents to these services will stay local. Moreover, many of these instances survive closer scrutiny.

4.2.1 Convergent Traceroutes. One explanation for the presence of convergent CitMs is that the Atlas probe is connected to a foreign-owned network, which may route domestic traffic through other countries. In our dataset, we find several instances of this happening. For example, we find probes in Mexico and Russia connected to the "ANEXIA Internetdienstleistungs GmbH" AS organization, which then peers with Arelion (a telecom based in Sweden), and so some traceroutes have Swedish hops. Another example is that a Brazil probe connected to the U.S.-owned "Neustar Security Services" network then sends traffic to U.S. hops in the U.S.-owned "Tata Communication" network. This case is particularly concerning, as the destination website is the email log-in page for the Brazilian Air Force, and Brazil has been vocal about combating U.S. surveillance of Internet traffic [27].

Another reason why convergent CitMs may occur is that domestic ASes may peer with foreign ASes. We again see many examples of this in our dataset. In one case, the Iraqi network "Zana Mohammed Mahdi A.Rahman company for Internet Service Provider LTD" peers with the U.S.-based network "Cogent Communications." In another example, we find a RIPE Atlas probe in Georgia, connected to a Georgian network[2] that routes traffic through Russia while heading to the Georgia government website for visas. Specifically, the probe's network "ZAO Aquafon-GSM" peers with the Russian AS "Dmitriy V. Kozmenko."

4.2.2 Divergent Traceroutes. Similarly, spot-checking a variety of divergent cases yield a number of frequent, plausible scenarios. Some are geographical, where adjacent nations appear on routes to foreign targets (e.g., South Africa's appearance as a CitM on paths originating in Tanzania) or island nations that make use of submarine cables landing at multiple CitMs on their way to the mainland. Others seem to be artifacts of colonial relationships, which have historically affected the development of submarine cables [63]. For example, Botswana has the U.K. as a CitM for all of its 16 divergent traceroutes. Similarly, Lebanon and Tunisia have

[2]The probe is connected to the Aquafon telecom, which is in Abkhazia, a partially-recognized state that most countries recognize as part of Georgia.

France as a CitM for over 90% of their divergent paths with CitMs.

4.3 False positives

On the other hand, the results of the pilot study contain a number of CitMs of which we are suspicious, based on geolocation information extracted from hostnames via the CAIDA Hoiho tool [34]. These examples include:

- The Czech Republic appears as a CitM in paths to nine different convergent websites for a single Japanese probe. These include sensitive websites such as a website to schedule COVID-19 vaccinations as well as subdomains of Japan's National Police Agency. However, upon further inspection, Hoiho geolocates all of these alleged CitM IP addresses to Japan, meaning these paths do not contain any actual CitMs.
- Similarly, a single Canadian probe reports seven paths to convergent websites as having the U.S. as a CitM. Several of these websites are related to the Canadian government's VPN tool [45]. While the U.S. is geographically close to Canada and thus plausible, Hoiho actually reports these IP addresses as residing in Canada. Again, this information eliminates all CitMs for these paths.
- Finally, two paths to the login portal for the Reserve Bank of Australia appear to have the U.S. as a CitM, but according to Hoiho, one of the IP addresses allegedly in the U.S. actually geolocates to Australia.

We additionally find one inaccuracy in the U.S. list of non-`.gov` domains: the website `nc-ddc.org` was added in 2013 as the website for the North Carolina Council on Developmental Disabilities. However, when visited in 2023, this is a Japanese website with information on what is safe and unsafe to feed dogs. We submitted a pull request to remove this domain from the list of official non-`.gov` domains, but it suggests that even government websites from official datasets may not be trustworthy and must be manually inspected.

5 FOCUSED STUDY

The anecdotes above are but a few of the dubious CitMs we discovered in our pilot study, causing us to refine our methodology to address a number of shortcomings that gave rise to many of the spurious datapoints. Some aspects of validation (validating government websites, geolocations, and traceroute responses) require manual analysis and corroboration with external sources in post processing. (Appendix C discusses how this post-processing impacts the conclusions we draw from our data set.) In the interest of reproducibility, we will make all artifacts available upon publication.

5.1 Country selection

Due to the extensive manual validation we perform to confirm the data, we limit the scope of our reported results to 11 countries, each of which host at least 20 active RIPE Atlas probes. Our selection criteria for countries is to maximize geographic diversity (including at least one country per continent) and diversity of Internet Freedom score [21], a measure of a country's digital rights (e.g., freedom from censorship, right to Internet access). For the latter criteria, the selected countries have Internet Freedom scores roughly following an even distribution with a slight skew towards countries with greater freedom.

5.2 Methodological refinements

To increase the number of seeds for our Censys search beyond the few sites listed in the UN website, we compile lists of official government websites for each country collected from various sources. We continue to use the official datasets for the United States [1, 46] as well as for Brazil [50]. For eight other countries, we find official, but much more limited, lists of websites for government organizations. For the remaining country we study, China, we were unable to identify any official source, so we use Wikipedia to find links to government agencies [65]. We detail the specific sources used for each country in Appendix B.

5.2.1 Website filtering. Unlike our pilot study, which only confirmed the existence of a webserver at target FQDNs, for our full study we deploy a VPN-hosted client within each country and manually inspect the returned webpage to ensure that it hosts content that belongs to the respective government. Concretely, to identify a target set of domains to probe, we manually verify whether each candidate URL hosts a government website to the best of our ability. Similar to prior work [31], we use ProtonVPN, a popular VPN with endpoints in almost all of the studied countries, to open a Web connection from an endpoint within the relevant country to the candidate URL. If there is no ProtonVPN endpoint in the selected country, we choose a ProtonVPN endpoint in a neighboring country to visit those websites (e.g. we use the Hong Kong VPN endpoint to visit Chinese websites). We acknowledge that this approach may result in discarding valid government websites that have strict geofencing or that block connections from VPNs.

If we are successful in contacting a webserver, we inspect the returned webpage (sometimes employing machine translation) to check if it belongs to an official government agency/body (both at federal and state/province levels, e.g. president and ministry websites), or a state-owned enterprise. We repeat this process until we obtain at least 100 *validated URLs* for each country, or exhaust our list of candidate URLs.

Country	URLs			Probed			Traceroutes
	Official	Certificates	Validated	Domains	IPs	Vantage Points	
Belarus	100	253	91	91	83	9	819
Brazil	524	614	102	100	123	10	989
Canada	121	630	100	100	115	10	998
China	40	171	101	99	253	9	885
Israel	229	225	101	99	105	8	792
Kenya	83	135	100	100	88	9	849
Philippines	62	116	109	99	144	10	984
Singapore	120	225	196	100	115	10	994
Taiwan	168	280	101	99	223	10	988
USA	18762	N/A	100	100	119	10	998
UAE	45	120	98	98	120	10	976

Table 1: For each country, we curate a list of potential target domains from official sources, expand the set using certificates, and manually validate each. We select up to 100 domains to probe from up to 10 vantage points resulting in almost 1,000 traceroutes before data sanitization.

We first examine URLs from the official sources and then randomly select among those obtained from our Censys search.

5.2.2 Post-processing. In the pilot study, we conservatively filter out countries that only appear once in a given traceroute in an effort to reduce spurious CitMs, but the heuristic is obviously imperfect. As an alternative, we considered an automated verification method that compared the latency between the purported CitM hops and Atlas probes located in both the source country and the CitM. Specifically, we collected `ping` measurements from up to five probes in both the source country and the alleged CitM and attempted to use Student t-tests to check for statistical differences in the ping latencies. Unfortunately, we find that the distribution of latencies from even a single probe may vary, and as previously mentioned, we cannot trust the self-reported geolocation of probes, mooting this approach. Another technique we considered was to cross-validate IPinfo's geolocation with that reported by the CAIDA Hoiho tool [35]. Unfortunately, Hoiho works based on hostnames, which are only available for 26% of the IP addresses in our dataset.

Instead, in our focused study, we validate the geolocation of each IP address in our traceroutes that corresponds to a potential CitM using a highly conservative RTT-based method from prior work [8, 28]. If the reported latency between a given hop and the source probe (as opposed to the prior hop [15]) is less than 2/3 the speed of light between the source and the hop's IPinfo-inferred country, we consider IPinfo to be erroneous. If the geolocation information for all of hops in a traceroute that were inferred to be located within a particular CitM are classified as errors using the aforementioned criteria, we *modify* the traceroute by removing those hops (i.e., discard the CitM). We do not discard the traceroute entirely, nor do we attempt to correct the geolocation—likely resulting in an under-reporting of CitMs for this type of traceroute.

6 DATA SANITIZATION

We collected the traceroutes used for our focused study in February 2024. The number of URLs, vantage points, and traceroutes collected per country is reported in Table 1. We manually inspect each of the 2,088 IP hops corresponding to CitMs in our dataset. This section details instances where we discard, or—when appropriate—modify traceroutes (by filtering out hops with invalid IP addresses). Inspection of the data reveals that one Brazilian probe (7113) returned empty responses for every traceroute so we discard all 93 paths collected by the probe. More fine-grained inspection discards or modifies an additional 411 traceroutes. We summarize the quantities in Table 2 and detail the reasons below.

6.1 Verifying Government Websites

Unfortunately, our manual process to validate government websites is not foolproof, especially when foreign languages are involved. Our initial set of 100 validated domains for Belarus includes `dha.by` which we extracted from a list of official Belarusian websites published by the Belarus Ministry of Foreign Affairs [47]. We discovered in subsequent analysis that this website does not, in fact, belong to the government. Hence, we discard paths to it in post-processing. We similarly discard paths to a Belarusian social media site that was similarly listed among its governmental websites.

Country	Total Traceroutes	Traceroutes Discarded			Traceroutes Modified		Subtotal
		No Info	Website	Probe	Geolocation	Squatting	
Belarus	819	0	18	0	0	0	18
Brazil	989	93	0	0	1	18	112
Canada	998	0	0	99	0	0	99
China	885	0	0	198	0	23	221
Israel	792	0	0	0	5	0	5
Kenya	849	0	0	0	0	0	0
Philippines	984	0	0	0	40	0	40
Singapore	994	0	0	0	0	0	0
Taiwan	988	0	0	0	0	0	0
USA	998	0	0	0	9	0	9
UAE	976	0	0	0	0	0	0

Table 2: The number of traceroutes discarded or modified per country. *No Info* indicates that an empty traceroute is returned; *Website* indicates the website is not an actual government website; *Probe* indicates we do not trust the self-reported geolocation for the probe; *Geolocation* indicates that some hops seem to be improperly geolocated; and *Squatting* breaks out traceroutes containing hops whose geolocation errors we suspect are due to the use of IP addresses commonly used for IP squatting. Cells with non-zero values are highlighted.

6.2 Filtering Vantage Points

A recent study reported that some RIPE Atlas probes misreport their own location [28]. In particular, by comparing RTTs from the probe's self-reported geolocation to servers with known geolocations, the authors identify violations of the 2/3 speed-of-light threshold. We find one Canadian probe (6493) in our dataset on their published list of improperly located Atlas probes and discard all traceroutes collected from this probe in our dataset. We also discover that two of the allegedly Chinese probes (7030 and 50179) we use are actually located in Hong Kong according to their self-reported latitude/longitude, which we analyze separately for the purposes of this study; hence, we discard paths from these probes.

6.3 IP squatting

Almost half of the improperly geolocated hops we discover in our dataset are due to IP squatting, so we call them out separately in Table 2. IP squatting refers to the use of IPv4 addresses that were historically allocated but not announced. We see 41 traceroutes in Brazil and China that experience CitMs for hops in this the 11.0.0.0/8 IP address range assigned to the U.S. DoD Network Information Center (AS749), which prior work found is often used for IP squatting [56]. Specifically, four China probes (all hosted within Alibaba) report traceroutes containing such DoD IP addresses—implying the U.S. as a CitM—for 25 traceroutes. We modify these 25 traceroutes by discarding the hops corresponding to IP addresses in this subnet. Similarly, 18 traceroutes from Brazilian vantage points also include hops in the 11.0.0.0/8 range; we again modify these traceroutes by discarding these hops. We may miss additional cases of IP squatting, especially for IP addresses announced by smaller organizations.

6.4 Interpreting Traceroute Responses

We explore the impact of two limitations of traceroute: unreliable responses and paths that do not reach the destination.

6.4.1 Unlabeled hops. Unlabeled hops in a traceroute (i.e., IP hops that do not generate ICMP responses) may hide additional CitMs. This means our study provides a lower bound on the number of CitMs experienced by each country. In order to get a sense of the scale of missing data, we plot a CDF of the longest set of consecutive unlabeled hops for each country in Figure 4. This figure presents two main takeaways. First, there is pervasive missing data: no more than 40% of any country's traceroutes are complete. Future work has the opportunity to incorporate information before and after unlabeled hops (as well as other information like BGP announcements) to infer what these hops may be, but this is out of scope for our work. Second, countries are associated with

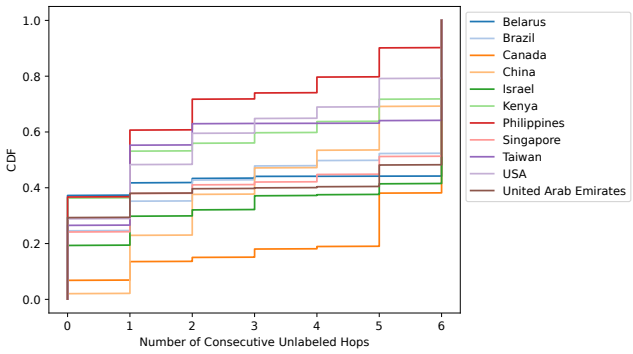


Figure 4: The CDF of the maximum number of consecutive unlabeled hops per traceroute for each of the countries we study.

different levels of transparency in the infrastructure they depend on to reach their government services. For example, the Philippines has the highest transparency in our dataset (61% of traceroutes do not have any consecutive unlabeled hops, i.e. values of 0 or 1 for the x-axis of Figure 4), while Canada has the lowest transparency (86% of traceroutes have consecutive unlabeled hops).

6.4.2 Traceroute reachability. Another limitation of traceroute is that many network devices filter out ICMP traffic, so some of our traceroutes are unable to reach the destination website. Only 57% of our traceroutes (5,888) reach the destination IP, with a very slight improvement to 58% of traceroutes (5,935) reaching the destination AS (i.e., the traceroute stops recording hops within the AS that originates the destination IP address). We do not exclude these results from our dataset because that could bias our data, e.g., we do not want to systematically exclude studying destinations hosted by Amazon (a hosting provider we find to frequently filter ICMP in our dataset), especially if we can still identify CitMs before the traffic is filtered.

That said, we acknowledge that the 3,929 traceroutes that do not reach the destination AS could under-report CitMs. As a way to characterize the portion of paths we miss, we attempt infer the missing ASes through which the traceroute could have traversed by comparing the set of ASes identified on the traceroute with AS paths in BGP advertisements. Using the February 2024 routing-table snapshots from Routeviews [51] and RIPE RIS [41] projects, we first remove BGP routes that had been observed for less than five days to avoid spurious paths. We extract the BGP routes whose AS paths contain both the source and destination ASes from at least one of our incomplete traceroutes. We are able to match 410 traceroutes (i.e., about 10% of the incomplete traceroutes) to a corresponding BGP route, identifying 66 additional ASes that our traceroutes could have traversed. The top-five ASes that did not respond to the most traceroutes are a cloud hosting provider (Amazon) followed by ASes run by governments (Canada) and Local ISPs in Singapore and Israel. While interesting, this BGP analysis is only circumstantial, so we do not include these findings in the results reported in subsequent sections.

In an effort to determine whether our results might be biased by the use of ICMP—as opposed to the transport protocols used by typical Web traffic—we ran a brief follow-up study in March 2025 to compare ICMP, TCP, and UDP traceroutes for two countries: Kenya and the U.S. For both countries, TCP traceroutes reach the destination IP address more frequently than ICMP (92% vs 80% for the U.S., and 87% vs 75% for Kenya)[3] or UDP (which only reached the destination 24% of the time in the U.S. and 28% for Kenya). Manual investigation reveals that when ICMP is filtered and stops short of the final few hops revealed by TCP, these last hops are often within the destination AS and do not reveal any additional CitMs.

Perhaps more interestingly, there are a few cases where ICMP and TCP traceroutes for the same source/destination pair are routed through different transit providers. Yet, even in those cases, we find that both protocols traverse the same set of CitMs (which are furthermore the same as those seen in our main study data collected in February 2024). Hence, while we acknowledge that TCP traceroutes may provide greater coverage—and potentially additional CitMs—we believe that the CitMs revealed through our ICMP traceroutes are salient. Future work could explore comprehensively comparing traceroutes from both protocols.

7 RESULTS

Table 3 presents the fully sanitized (i.e., validated, post-processed, and manually inspected) results of our focused measurement study, showing the percentage of convergent and divergent traceroutes for each country as well as the prevalence of CitMs. Figure 5 visualizes these results in the style of Figure 3, with the addition of outlines surrounding each stacked bar that show the total volume of traceroutes collected, including those that were discarded during post-processing and/or sanitization (i.e., appear in Table 2). In this section we describe the high-level takeaways we extract from our results, as well as the negligible impact anycast seems to have on our findings. Additional anecdotes we discovered as part of our manual inspection are included in Appendix C.2.

7.1 Insights Gained

We extract three high-level trends that may explain the patterns of convergence/divergence and CitMs in our dataset.

[3]In 2024, 78% of paths from the U.S. and 70% of paths from Kenya reached. Other countries had lower reachability rates, lowering the average for 2024.

Country	Total Traceroutes	Non-Anycast Traceroutes	Convergent (%)	Divergent (%)	% CitM of		
					Overall	Convergent	Divergent
Belarus	801	801	98.88	1.12	0.00	0.00	0.00
Brazil	896	824	96.36	3.64	8.13	7.30	30.00
Canada	899	791	91.78	8.22	22.63	21.49	35.38
China	687	687	98.69	1.31	0.58	0.59	0.00
Israel	792	606	73.60	26.40	14.69	0.45	54.38
Kenya	849	791	63.66	36.34	37.97	10.83	85.52
Philippines	984	636	21.07	78.93	43.08	34.33	45.42
Singapore	994	735	84.22	15.78	15.24	14.22	20.69
Taiwan	988	938	94.46	5.54	15.57	11.96	76.92
USA	998	719	100.00	0.00	0.14	0.14	0.00
UAE	976	672	91.96	8.04	33.18	31.23	55.56

Table 3: Percent of traceroutes to convergent and divergent destinations *after data sanitization*, and percent of traceroutes with CitMs. Percentages are relative to the number of non-anycast traceroutes.

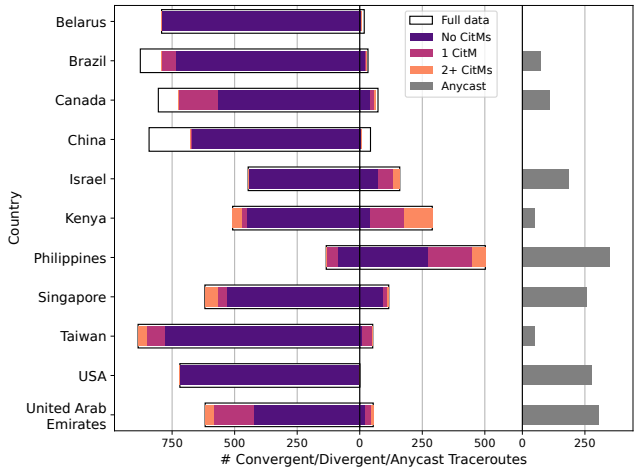


Figure 5: Frequency of CitMs in each country's results after data sanitization. Enclosing rectangles show the number of paths for each country before sanitization.

We emphasize that these are trends we observe, but each trend has some counter-examples.

(1) Countries with well-developed Internet infrastructure are more likely to host their websites domestically, but this relationship is not perfect. In fact, countries can take measures to increase their levels of convergence.
(2) The level of convergence vs divergence influences the amount of CitMs a country experiences, as divergent traceroutes may necessarily transit CitMs.
(3) Countries with higher levels of Internet infrastructure experience fewer CitMs.

7.1.1 Rate of convergence. We find meaningful variation in domestic vs. foreign hosting for each country, and find that countries that host more websites locally are often—but not always—those with robust Internet infrastructure. Seven of the 11 countries have strong domestic hosting infrastructure, as more than 90% of traceroutes reach convergent targets. Many of these countries (e.g. the U.S., Taiwan, China, Canada, and the U.A.E.) all have robust Internet infrastructure. However, one of these countries, Belarus, is not known for having a robust Internet infrastructure. We believe that the explanation for almost 99% of paths reaching convergent targets is a presidential decree that ordered all .by domains must be physically hosted within the country [22].

Similarly, countries with less developed Internet infrastructure often have more divergent targets. Countries like the Philippines and Kenya are still developing their infrastructure, and both countries have a high fraction of traceroutes to divergent websites. However, both Israel and Singapore have a surprisingly high fraction of traceroutes to divergent destinations given their robust infrastructure. Both countries host many of their government websites in the U.S., and Israel also hosts websites in Europe.

7.1.2 Relating CitMs to Geographic Factors. Intuitively, one may expect that paths to convergent destinations experience fewer CitMs than paths to divergent destinations, which may transit other countries out of geographic necessity. We observe some correlation between divergence and CitMs in our dataset. Israel, Singapore, and Canada all experience many European CitMs en route to web servers hosted in Europe. Brazil experiences nine traceroutes with the U.S. as a CitM en route to Canada. On the other hand, we see a number of CitMs on paths to convergent destinations—some of which may be explained by geographic proximity. The Philippines has the highest rate of CitMs for convergent targets at 43%; the most common CitM for these paths is Singapore, followed by Hong Kong. We now present two examples (with validation details) of geographically-proximal CitMs to convergent websites.

Hong Kong as a CitM for Taiwan. Two of our Taiwan probes are hosted by PCCW Global, but one consistently experiences Hong Kong as a CitM, and the other does not.[4] We find that almost 75% of the traceroutes for one probe visit Hong Kong on the second hop before returning to Taiwan on the third hop (via an IP address announced by Chunghwa Telecom). Meanwhile, the other probe also visits an IP address announced by Chunghwa Telecom in the second or third hop in almost 90% of its traceroutes, but never experiences a CitM.

For the probe experiencing the CitMs, 73 of the 99 traceroutes visit Hong Kong on the second hop and return to Taiwan on the third; on average, these traceroutes jump from a latency of 1.11 ms on the first hop to 47.25 ms on the second hop and 106.99 ms on the third hop. For its 19 traceroutes without any CitMs, on average the first three hops have latencies of 1.12 ms, 6.35 ms, and 69.76 ms—each far lower than the previous set of latencies, which supports the geolocation inferences for Hong Kong.

In the case of the probe that does not experience Hong Kong as a CitM, 86 of its 99 traceroutes visit Chunghwa Telecom (the same AS that the first probe reaches on the third hop) on the second or third hop. On average, RTTs of the second and third hops are 0.77 ms and 1.62 ms. Again, these are far lower than the latencies experienced in the other probe's paths, which suggests that Hong Kong is a plausible CitM and not a geolocation error.

Hong Kong as a CitM for Singapore. Of Singapore's 88 traceroutes to convergent websites with CitMs, almost half (38) see Hong Kong as a CitM. We find that 15 of these paths transit the Hong Kong Internet Exchange (HKIX), all from the same probe that likely relies on HKIX for connectivity. We find another probe also consistently sees Hong Kong as a CitM for its fourth hop and traceroute latencies spike from (on average) 1.96 ms on the third hop to 29.09 ms on the fourth hop, indicating the CitM is plausible.

7.1.3 CitMs vs. Internet Infrastructure. Intuitively, countries with well-developed economies and Internet infrastructure are more likely to be CitMs for many other countries. Figure 6 presents a heatmap to illustrate the frequency of CitMs to both convergent and divergent (i.e., not anycast) destinations. France, Germany, Hong Kong, Singapore, and the U.S. are common CitMs across countries around the world. Hong Kong is a frequent CitM for countries in Southeast Asia like the Philippines, Taiwan, and Singapore (as described above).

The heatmap also reveals strong relationships between certain pairs of countries. For example, Singapore is a common CitM for Indonesia, and the U.S. is a common CitM for Canada, both of which may be expected due to geographic proximity. However, it also reveals some unexpected pairings. For example, we see France and the U.K. as common CitMs for Kenya despite being geographically distant. We also see Singapore and France as common CitMs for the United Arab Emirates. We describe two case studies of CitMs for Kenya and the U.A.E. in this section. While Singapore has many CitMs in Africa and Europe, some of these are due to one probe (7219) hosted by Angola Cables, which sends all but one of its paths through South Africa; these CitMs likely do not generalize to residents as Angola Cables does not advertise services in Singapore.

U.K. and South Africa as CitMs for Kenya. A decade ago, prior work found that Liquid Telecom routed traffic from South Africa to Kenya through London [19]. Interestingly, as Kenya has developed its infrastructure, its traffic that transits Liquid Telecom often avoids CitMs. We observe two convergent destinations for which most probes (seven of eight) transit Liquid Telecom and do not experience CitMs, but one probe does experience CitMs while transiting Liquid Telecom.[5] The latter probe goes through the U.K. and South Africa before returning to Kenya for both destinations.

We validated these geolocation results through Hoiho and latency analysis. Hoiho confirmed one of the four U.K. addresses, but did not have information about the three South African ones. Another U.K. IP address maps to Liquid Telecom's IP address in the London Internet Exchange (LINX), supporting this inference. The latency also jumps from Kenya to the U.K. in both traceroutes (0.54ms to 194.51ms in one traceroute and 0.81ms to 197.21ms in the other).

[4]Probe 6181 sees CitMs, and probe 1002754 does not. Probe 6818's IP address is announced by both PCCW Global (AS3491) and Gateway Communications (AS31713). Gateway was acquired by PCCW in 2012.

[5]Probe 13218 experiences these CitMs for RIPE Atlas measurement IDs 68064916 and 68064948. One Kenyan probe malfunctions for these destinations, so only eight of the nine probes reported results.

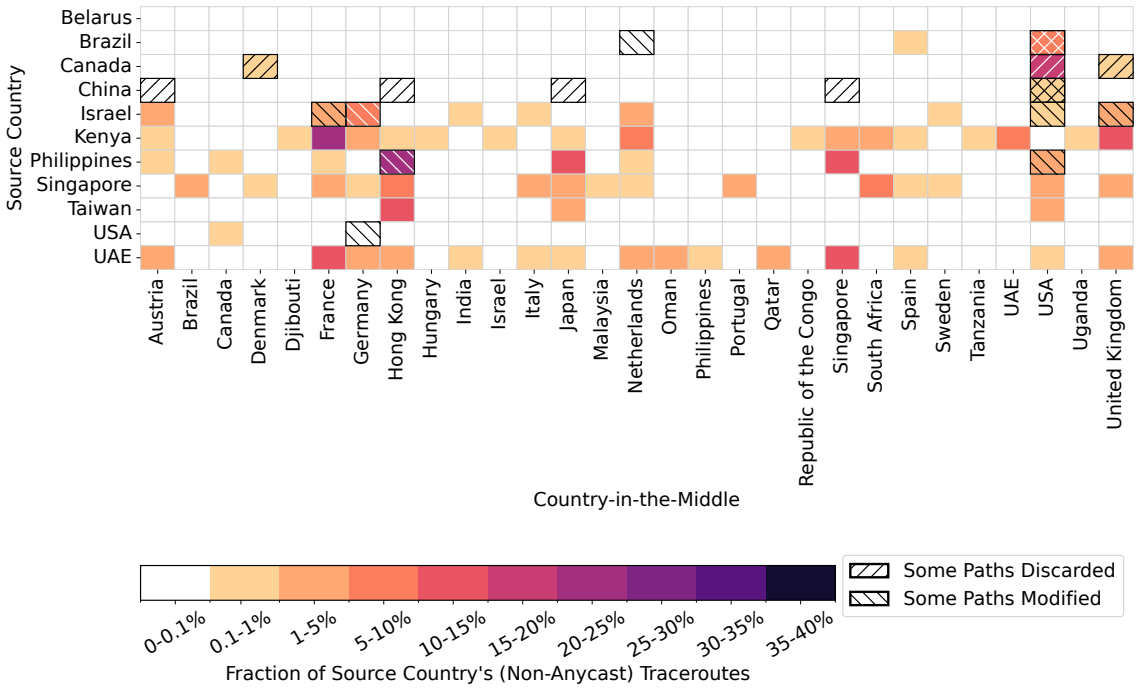


Figure 6: Heatmap of the CitMs for convergent and divergent targets after data sanitization. Hatched cells indicate traceroutes being discarded or modified from the full dataset (Section 6).

Singapore as a CitM for the U.A.E.. Singapore is a CitM for 14% of all paths to convergent destinations for the U.A.E. While Singapore is geographically far from the U.A.E., we cross-reference the 88 IP addresses where Singapore is a CitM with Hoiho and find agreement for 59 addresses (67%), indicating that Singapore is actually a CitM and not likely to be a geolocation error. We believe that Singapore actually makes sense because both countries are major data center hubs and are connected by multiple submarine cables (e.g. SEA-ME-WE 5 [58]).

7.2 Anycast

As in our pilot study, we compare the IP addresses of the target websites against the Anycast Census dataset [20], using the data collected closest in time to our traceroute collection (March 21, 2024) and find that 176 targets appear to be anycast (63.6% of which are hosted by Cloudflare). For traceroutes to these targets, we identify *anycast-CitMs*: countries on the path that are neither the source country nor any of the potential destination countries reported by the dataset. We find that 23% (395) of the 1,685 traceroutes to anycast destinations have at least one anycast-CitM. However, in all but one case, these anycast-CitMs match the existing CitMs seen in the non-anycast traceroutes for that country. There is one case where a probe from the U.A.E. supposedly directly connects to hops in Thailand before reaching the destination IP address, which the Anycast Census reports is hosted in Australia, Japan, India, Indonesia, or South Korea. Unfortunately, latency analysis is inconclusive given the close geographic proximity of Thailand to several of the potential host countries and we are unable to corroborate the purported Thai hops with other data sources such as Hoiho [34] so are uncertain whether it represents an additional CitM or not; otherwise the set of CitMs for anycast and unicast destinations in our dataset are the same.

8 RELATED WORK

Our work bridges two groups of prior work: analyzing the foreign dependencies of government websites [3, 23, 24, 30, 31, 60, 62], and analyzing CitMs [6, 13, 14, 18, 19, 59]. To the best of our knowledge, our study is the first to develop a method to identify and validate CitMs, and apply it to government websites.

8.1 Government Website Dependencies

Our work complements the growing body of literature that analyzed foreign dependencies of government websites' DNS infrastructure [3, 23, 62], certificate authorities [24, 60], hosting providers [3, 30, 31], and content providers [24]. One study of hosting providers [31] uses a methodology similar

Paper	Validates Hop Geolocation	Validates VP Geolocation	Considers Unreliable Responses	Considers Unreachable Traceroutes	Considers Anycast Sites
Gupta 2014 [19]	No	Yes	No	No	No
Fanou 2015 [14]	Yes	Yes	Yes	No	N/A
Shah 2016 [59]	No	No	Yes	No	No
Edmundson 2018 [13]	No	No	No	No	No
Gueye 2018 [18]	No	No	No	No	No
Candela 2021 [6]	No	No	No	Yes	No
Current Work	Yes	Yes	Yes	Yes	Yes

Table 4: Comparison of how prior work addresses the challenges we identify in our paper (excluding validating government websites, as that is not applicable for these works).

to ours. A 2022 study finds Russian domains are almost entirely convergent [30], similarly to a 2023 study that finds some government websites for Brazil, India, and South Africa are mostly convergent [3]. Finally, a 2019 study found that all G7 countries have some government websites that depend on foreign content providers [24].

8.2 Identifying CitMs

While we are not the first to attempt to measure the prevalence of CitMs (of which convergent paths with CitMs are often called "tromboning" or "boomerang" paths), most prior studies are more general in the paths they consider and lack rigorous validation. We consider whether prior work [6, 13, 14, 18, 19, 59] addresses four of the five challenges we identify (the challenge of validating government websites is not applicable): validating geolocation of 1) hops and 2) vantage points, considering 3) traceroute reachability and unreliability, and 4) anycast destinations. We summarize each paper's ability to navigate these challenges in Table 4.

Only one of these studies attempts to validate the geolocation of intermediate hops: Fanou *et al.* [14] cross-validates multiple geolocation datasets with each other and with ping latencies. Their study, as well as one by Gupta *et al.* [19], both address the challenge of validating geolocation of vantage points by controlling the deployment of these vantage points. Fanou *et al.* and Shah *et al.* [59] consider the unreliability of traceroute responses; the former notes the impact of unlabeled hops on traceroute latencies but does not separate or discard these traceroutes, while the latter discards traceroutes with fewer than three hops responding. Only Candela *et al.* [6] considers traceroute reachability, by discarding all traceroutes that do not reach the destination IP address; however, we believe this may bias results (Section 6.4.2). Finally, none of these studies explain how anycast destinations complicate their definitions of boomerang routes or detours. While our work does not completely solve all of these challenges, to the best of our knowledge, it is the first to consider the impact of each of these challenges on our findings.

9 RECOMMENDATIONS

Traffic sovereignty has become increasingly important for governments that wish to understand how their network traffic flows through other countries. However, it is extremely difficult to understand these dependencies for a large number of countries due to five core challenges: IP geolocation, collecting government websites, finding appropriate vantage points, interpreting traceroute responses, and handling anycast websites. Our work approaches these challenges through corroboration with external sources and extensive manual validation with some success, but we do not solve all of these challenges. As the research community continues to tackle these challenges, our framework will also improve.

Our work leads to several recommendations for measurement researchers and opportunities for future work. Our methodology relies on extensive manual effort (e.g. visiting government websites, identifying geolocation errors) to validate our dataset. Automating these manual efforts could enable larger-scale analyses of CitMs. We recommend that researchers use extra caution when conducting measurements from third-party vantage points, as self-reported geolocation may be incorrect (Section 6.2). We also recommend that the research community continue to work on improving geolocation accuracy for the Internet core and identifying IP squatting. For identifying CitMs, promising areas of future work include inferring CitMs from unlabeled hops. Future work could also investigate the presence of CitMs in paths for TLS handshakes. Finally, more targeted, country-level studies of CitMs can sidestep many challenges of a global study (e.g., by deploying vantage points with known geolocations).

REFERENCES

[1] U.S. Cybersecurity & Infrastructure Security Agency. .gov data. https://github.com/cisagov/dotgov-data.

[2] NTT America. Network Map.

[3] Demétrio F Boeira, Eder J Scheid, Muriel F Franco, Luciano Zembruzki, and Lisandro Z Granville. Traffic Centralization and Digital Sovereignty: An Analysis Under the Lens of DNS Servers. *arXiv preprint arXiv:2307.01300*, 2023.

[4] UAE Cabinet. Category:government ministries of the uae. https://uaecabinet.ae/en/ministries-and-federal-authorities.

[5] CAIDA. Routeviews Prefix to AS mappings Dataset (pfx2as) for IPv4 and IPv6, 2022.

[6] Massimo Candela, Valerio Luconi, and Alessio Vecchio. A worldwide study on the geographic locality of internet routes. *Computer Networks*, 2021.

[7] Michael Dalton et al. Andromeda: Performance, Isolation, and Velocity at Scale in Cloud Network Virtualization. In *Proc. of 15th USENIX NSDI*, 2018.

[8] Omar Darwich, Hugo Rimlinger, Milo Dreyfus, Matthieu Gouel, and Kevin Vermeulen. Replication: Towards a Publicly Available Internet Scale IP Geolocation Dataset. In *Proc. of 2023 ACM IMC*, 2023.

[9] Benoit Donnet, Matthew Luckie, Pascal Mérindol, and Jean-Jacques Pansiot. Revealing MPLS tunnels obscured from traceroute. *ACM SIGCOMM Computer Communication Review*, 2012.

[10] Ben Du, Massimo Candela, Bradley Huffaker, Alex C. Snoeren, and KC Claffy. RIPE IPmap Active Geolocation: Mechanism and Performance Evaluation. *SIGCOMM Comput. Commun. Rev.*, May 2020.

[11] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. A Search Engine Backed by Internet-Wide Scanning. In *Proc of. 22nd ACM CCS*, 2015.

[12] Zakir Durumeric, Eric Wustrow, and J Alex Halderman. ZMap: Fast Internet-wide Scanning and Its Security Applications. In *USENIX Security Symposium*, volume 8, 2013.

[13] Anne Edmundson, Roya Ensafi, Nick Feamster, and Jennifer Rexford. Nation-State Hegemony in Internet Routing. In *Proc. of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018.

[14] Rodérick Fanou, Pierre Francois, and Emile Aben. On the Diversity of Interdomain Routing in Africa. In *Proc. of 16th PAM*, 2015.

[15] Romain Fontugne, Cristel Pelsser, Emile Aben, and Randy Bush. Pinpointing Delay and Forwarding Anomalies Using Large-Scale Traceroute Measurements. In *Proc. of 2017 ACM IMC*, 2017.

[16] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. A Look at Router Geolocation in Public and Commercial Databases. In *Proc. of ACM IMC*, 2017.

[17] Matthieu Gouel, Kevin Vermeulen, Olivier Fourmaux, Timur Friedman, and Robert Beverly. IP Geolocation Database Stability and Implications for Network Research. In *Proc. of IEEE/IFIP TMA*, 2021.

[18] Assane Gueye and Babacar Mbaye. On the Prevalence of Boomerang Routing in Africa: Analysis and Potential Solutions. In *Proc. of InterSol*, 2018.

[19] Arpit Gupta, Matt Calder, Nick Feamster, Marshini Chetty, Enrico Calandro, and Ethan Katz-Bassett. Peering at the Internet's Frontier: A First Look at ISP Interconnectivity in Africa. In *Proc. of 15th PAM*, 2014.

[20] Remi Hendriks, Matthew Luckie, Mattijs Jonker, Raffaele Sommese, and Roland van Rijswijk-Deij. MAnycast Reloaded: a Tool for an Open, Fast, Responsible and Efficient Daily Anycast Census. *arXiv preprint arXiv:2503.20554*, 2025.

[21] Freedom House. Internet Freedom Scores. Accessed 2024-10-04. https://freedomhouse.org/countries/freedom-net/scores.

[22] Freedom House. Belarus: Freedom on the Net 2022, 2022.

[23] Rebekah Houser, Shuai Hao, Chase Cotton, and Haining Wang. A Comprehensive, Longitudinal Study of Government DNS Deployment at Global Scale. In *Proc. of 52nd IEEE/IFIP DSN*, 2022.

[24] Hsu-Chun Hsiao, Tiffany Hyun-Jin Kim, Yu-Ming Ku, Chun-Ming Chang, Hung-Fang Chen, Yu-Jen Chen, Chun-Wen Wang, and Wei Jeng. An Investigation of Cyber Autonomy on Government Websites. In *Proc. of ACM Web Conf.*, 2019.

[25] Amnesty International. Belarus uses telecoms firms to stifle dissent, July 2016.

[26] ipinfo.io. IP geolocation API.

[27] Esteban Israel and Anthony Boadle. Brazil conference will plot internet's future post nsa spying. *Reuters*, April 2014. https://www.reuters.com/article/us-internet-conference-idUSBREA3L1OJ20140422.

[28] Katherine Izhikevich, Ben Du, Sumanth Rao, Alisha Ukani, and Liz Izhikevich. Trust, But Verify, Operator-Reported Geolocation. *arXiv preprint arXiv:2409.19109*, 2024.

[29] Liz Izhikevich, Gautam Akiwate, Briana Berger, Spencer Drakontaidis, Anna Ascheman, Paul Pearce, David Adrian, and Zakir Durumeric. ZDNS: A Fast DNS Toolkit for Internet Measurement. In *Proc. of 2022 ACM IMC*, IMC '22, 2022.

[30] Mattijs Jonker, Gautam Akiwate, Antonia Affinito, KC Claffy, Alessio Botta, Geoffrey M Voelker, Roland van Rijswijk-Deij, and Stefan Savage. Where .ru? Assessing the Impact of Conflict on Russian Domain Infrastructure. In *Proc. of 22nd ACM IMC*, 2022.

[31] Rashna Kumar, Esteban Carisimo, Lukas De Angelis Riva, Mauricio Buzzone, Fabián E. Bustamante, Ihsan Ayyub Qazi, and Mariano G. Beiro. Of Choices and Control - A Comparative Analysis of Government Hosting. In *Proc. of ACM IMC*, 2024.

[32] Matthew Luckie and KC Claffy. A Second Look at Detecting Third-Party Addresses in Traceroute Traces with the IP Timestamp Option. In *In Proc. of PAM*, 2014.

[33] Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, David Clark, and kc Claffy. bdrmap: Inference of borders between IP networks. In *Proc. of 2016 ACM IMC*, 2016.

[34] Matthew Luckie, Bradley Huffaker, and KC Claffy. Learning Regexes to Extract Router Names from Hostnames. In *Proc. of ACM IMC*, 2019.

[35] Matthew Luckie, Bradley Huffaker, Alexander Marder, Zachary Bischof, Marianne Fletcher, and KC Claffy. Learning to Extract Geographic Information from Internet Router Hostnames. In *Proc. of 17th CoNEXT*, 2021.

[36] Ewen MacAskill, Julian Borger, Nick Hopkins, Nick Davies, and James Ball. GCHQ taps fibre-optic cables for secret access to world's communications. *The Guardian*, June 2013.

[37] MaxMind: IP Geolocation and Online Fraud Prevention, 2024.

[38] Republic of China (Taiwan) Ministry of Foreign Affairs. Government agencies. https://www.taiwan.gov.tw/3866.php?xq_xcat=5.

[39] United Nations. E-government knowledgebase country data. https://publicadministration.un.org/egovkb/en-us/Resources/Country-URLs.

[40] United Nations. Non-self-governing territories, 2022. https://www.un.org/dppa/decolonization/en/nsgt.

[41] RIPE NCC. Routing Information System (RIS), 2024.

[42] NetAcuity. https://digitalelement.com/solutions/ip-location-targeting/netacuity.

[43] Jonathan A Obar and Andrew Clement. Internet Surveillance and Boomerang Routing: A Call for Canadian Network Sovereignty. In *Proc. of TEM 2013*, 2012.

[44] Government of Canada. Departments and agencies. https://www.canada.ca/en/government/dept.html.

[45] Government of Canada. Making the right connections with VPN. Accessed 2025-05-15. https://www.canada.ca/en/shared-services/

campaigns/stories/right-connections.html.
[46] U.S. Department of Defense. Dod websites. https://www.defense.gov/Resources/Military-Departments/DOD-Websites/.
[47] Ministry of Foreign Affairs of the Republic of Belarus. Links to official belarusian web-sites. https://mfa.gov.by/en/links/.
[48] Government of Israel. Government ministries of israel. https://www.gov.il/en/departments.
[49] Government Procurement Administration of Israel. Governmental and public organizations. https://mr.gov.il/ilgstorefront/en/gov-websites.
[50] Brazil Ministry of Management and Innovation in Public Services. Data from the organizational structure of the federal executive branch (siorg system). https://dados.gov.br/dados/conjuntos-dados/dados-da-estrutura-organizacional-do-poder-executivo-federal-sistema-siorg.
[51] University of Oregon. Route Views Project, 2024.
[52] Government of Singapore. Trusted sites. https://www.gov.sg/trusted-sites.
[53] Office of the Deputy President. Category:government ministries of kenya. https://www.devolution.go.ke/county-websites/.
[54] Republic of the Philippines. List of government websites. https://www.officialgazette.gov.ph/lists/government-websites/.
[55] Office of the President. Category:government ministries of kenya. https://www.president.go.ke/ministries-ke/.
[56] Loqman Salamatian, Todd Arnold, Ítalo Cunha, Jiangchen Zhu, Yunfan Zhang, Ethan Katz-Bassett, and Matt Calder. Who Squats IPv4 Addresses? *ACM SIGCOMM Computer Communication Review*, 2023.
[57] Adam Satariano and Scott Reinhard. How russia took over ukraine's internet in occupied territories. *The New York Times*, 08 2022.
[58] SEA-ME-WE 5 Maps.
[59] Anant Shah, Romain Fontugne, and Christos Papadopoulos. Towards Characterizing International Routing Detours. In *Proc. of the 12th AINTEC*, 2016.
[60] Sudheesh Singanamalla, Esther Han Beol Jang, Richard Anderson, Tadayoshi Kohno, and Kurtis Heimerl. Accept the Risk and Continue: Measuring the Long Tail of Government https Adoption. In *Proc. of ACM IMC*, 2020.
[61] Raffaele Sommese, Leandro Bertholdo, Gautam Akiwate, Mattijs Jonker, Roland van Rijswijk-Deij, Alberto Dainotti, KC Claffy, and Anna Sperotto. MAnycast2: Using Anycast to Measure Anycast. In *Proc. of ACM IMC*, IMC '20, 2020.
[62] Raffaele Sommese, Mattijs Jonker, Jeroen van der Ham, and Giovane CM Moura. Assessing e-Government DNS Resilience. In *Proc. of 18th IEEE CNSM*, 2022.
[63] Nicole Starosielski. *The Undersea Network*. Duke University Press, 2015.
[64] New Zealand The Philippine Embassy of Wellington. Ph government websites. https://www.officialgazette.gov.ph/lists/government-websites/.
[65] Wikipedia. Category:government agencies of china. https://en.wikipedia.org/wiki/Category:Government_agencies_of_China.

A ETHICS

We attempted to reduce the burden of our traceroutes to the domains we probe by using at most 10 RIPE Atlas probes.

B COLLECTION PROCESS FOR OFFICIAL GOVERNMENT WEBSITES

In this section we describe how we collected official government websites for each country in our focused study:

Belarus. The Belarus Ministry of Foreign Affairs publishes a list of 145 official government websites [47]. We visited this website on January 30, 2024; it is unknown when the website was last updated.

Brazil. The Brazilian Ministry of Management and Innovation in Public Services publishes a dataset called "Organizational Structure of the Federal Executive Branch" [50]. The dataset is updated monthly and we used the latest available dataset, which was created on January 2, 2024. The dataset contains 1,494 unique URLs.

Canada. The Canadian government publishes a website with its departments and agencies [44], which was last updated on October 4, 2023. We scraped the links on this website on January 24, 2024. Some of these links were to internal pages; in this case, we first visited the internal link and checked if it redirected to an external website. If not, we collected all internal links in a `main` HTML element. This process resulted in 138 unique URLs.

China. We used Wikipedia to find a list of government agencies belonging to China, with urls linking to another Wikipedia page about each agency [65]. On January 23, 2024 we manually opened each Wikipedia page and looked for either an "External Links" section that linked to the ageny's own website or we found the website linked in the "Agency Overview" under the "Website" section. This process resulted in 42 unique URLs.

Israel. The Israeli government publishes a website with its government ministries [48] and the Government Procurement Administration publishes a list of government websites [49]. We scraped the links on January 22, 2024. This process resulted in 309 unique URLs.

Kenya. The Kenyan Office of the President [55] provides a list of all national ministries. Their Office of the Deputy President [53] provides a list of 47 county government websites. We scraped those 2 lists on Jan 24, 2024 and obtained 89 unique URLs.

Philippines. The Philippines government publishes a list of official government websites for some of its federal organizations [54]. We supplemented this list with a list of government websites from the Philippines Embassy of New Zealand [64]. We visited both websites on August 28, 2023 and collected a total of 79 unique URLs.

Singapore. The Singapore government claims that most of their government websites have the form `*.gov.sg`; however, they publish a list of 125 official government websites that do not follow this pattern [52]. We visited this website on January 24, 2024. It is unknown when this website was last updated.

Taiwan. The Taiwanese government publishes a list of government agencies on their website [38]. We visited this website on January 26, 2024 and collected a total of 295 unique URLs.

United Arab Emirates. The Cabinet of the UAE provides a list of ministries and federal authorities [4]. We scraped the website on Jan 24, 2024 and obtained 46 unique URLs.

USA. As mentioned in Section 3, we combine three official data sources. The first source is a set of URLs ending in `.gov`, which is updated every day [1] and contains 9,833 unique URLs. The second source is a set of non-`.gov` domains; we use the `1_govt_urls_full.csv` file, which contains domains across federal, state, regional, county, and local levels, as well as native sovereign nations (i.e., tribal nations) and quasigovernmental organizations. This file contains 9,226 unique URLs. The third and final source is a set of U.S. military domains from the Department of Defense website [46], from which we exclude social media websites for a total of 593 unique URLs.

In total, these data sources contained 19,299 unique URLs. This number is more than the sum of the URLs in each data source as there were some duplicates between the military URLs and the other two datasets.

Each of these sources were visited on January 30, 2024; at the time we visited them, the `.gov` dataset was last updated that same day, and the non-`.gov` domains were last updated on January 30, 2023 (a year prior). It is unknown when the website with military domains was last updated.

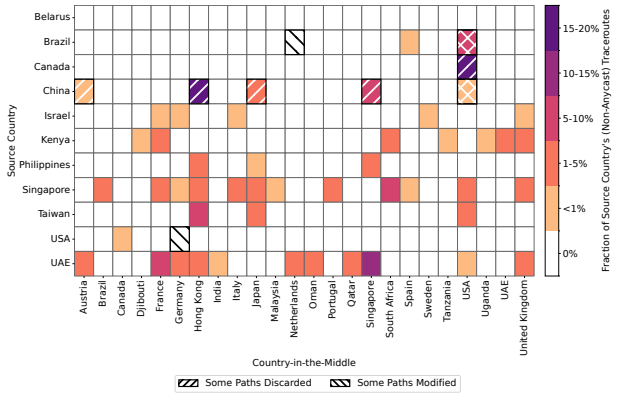
C ADDITIONAL VALIDATION DETAILS

We use Censys [11] to expand our set of government domains collected from authoritative sources (Section 3.1.2). This process identifies a large number of URLs in each country, but some of these may be unreachable, not serve webpages, or may not belong to the government. We considered filtering to only "trusted certificates," i.e. certificates where the certificate chain can be followed to a root certificate found in a major root trust store. However, using only trusted certificates filters out legitimate government websites. For example, 23 legitimate Belarussian governmental websites (ending in `.gov.by`) use untrusted TLS certificates. Instead, we filter the Censys results to subdomains of the domains and eTLD+1's that we obtained through authoritative sources (Section 3.1.2) and manually inspect each website.

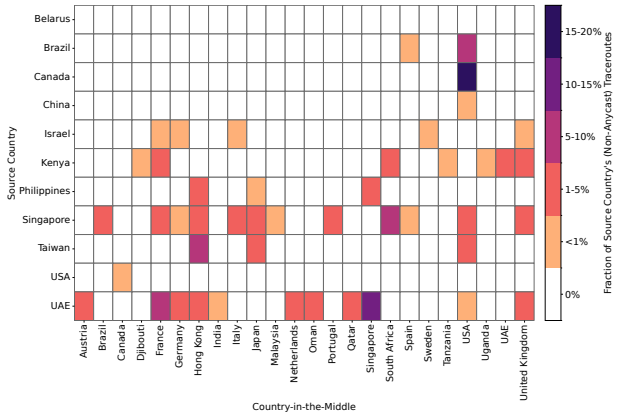
C.1 Impact of Validation

In Figure 7 we present a heatmap of the paths to convergent websites that contain CitMs, for both the full and the refined datasets (see Figure 6 for a similar heatmap for both convergent and divergent destinations). After data sanitization, all but one CitM disappear for China, as most of the CitMs were experienced by the probes we exclude.

We find relatively few geolocation errors overall because 1) we use a high threshold of speed-of-light violation for identifying errors, and 2) we focus our efforts on confirming



(a) Full (unrefined) dataset



(b) Sanitized dataset

Figure 7: Heatmap of CitMs on convergent paths (a) full and (b) sanitized datasets. Cross-hatching indicates traceroutes that were either discarded or modified from the full dataset.

CitMs identified by the geolocation dataset we use; we do not try to find additional CitMs that were not identified by IPinfo.

C.2 Interesting anecdotes

This subsection contains a variety of potentially interesting CitMs we uncovered during our manual validation.

C.2.1 CitM Path due to Telxius Cables. We find two traceroutes to the destination website `csjt.jus.br` (for Brazil's Superior Council of Labor Justice) that transit Spain and the U.S. en route to the destination website in Colombia (hosted by Amazon). Both paths traverse Telefónica Global Solutions, and within this network the paths are first routed to the U.S.

and then Spain; after these hops, we see four unlabeled responses before reaching a Colombia IP address associated with Amazon.

Based on the publicly available network maps from Telxius (a child company of Telefónica), we see that this path maps onto high-capacity submarine cables: the Firminia or Brusa cables that connect Brazil to the U.S., and the Marea cable that connects the U.S. to Spain [2]. However, this map does not show a direct connection from Spain to Colombia. If the path were to continue within the Telxius network and the public network map, the path would need to go back to the U.S., go from Virginia Beach to Ashburn via terrestrial backhaul, go to Jacksonville via an extended fiber route, and then go to Colombia via a submarine cable. Interestingly, the Telxius SAm-1 cable directly connects Brazil and Colombia and thus offers a path without CitMs; however, we do not see this route (or any other IP addresses for Colombia) in our data.

Hoiho does not contain any validation information for this traceroute. Latencies spike from around 7ms to 130ms for the US hop and then stay around 130ms for the hops to Spain and Colombia.

C.2.2 US as CitM for China. We find two traceroutes from a single probe (53274) that have the U.S. as a CitM. For this probe's paths without CitMs, the third hop is always an IP address announced by No.31,Jin-rong Street (AS4134). However, in the paths with CitMs, hops 3 and 4 remain in the 192.0.0.0/8 space, and the fifth hop geolocates to the U.S. and is announced by EGIHosting (AS18779). The latency for the fifth hop indicates that visiting the U.S. is plausible, but we are unable to further confirm this behavior.

C.2.3 Canada as CitM for US. A Seattle probe produces one path with one CitM hop, which geolocates to Vancouver. The IP address, as well as the ones immediately before and after in the path, are within the AT&T network. We are unable to corroborate the geolocation, as Hoiho does not have geolocation information for any of these AT&T IP addresses. However, IPinfo indicates that the hops preceding the CitM remain in Seattle and Portland, indicating that this path is plausible.

C.2.4 CitMs for Canada. The U.S. is the most prevalent CitM for Canada. In particular, we find a common pattern for the 168 paths with CitMs to convergent targets: the source RIPE Atlas probes are located in Vancouver, and the destination is hosted in cities in eastern Canada such as Ottowa and Montreal. The CitM hops traverse mostly northern U.S. cities such as Seattle, Chicago, New York, but also some southern cities like Virginia and Dallas. The CitM hops are announced by ASNs for major U.S. cloud providers like Microsoft and Amazon, as well as Tier-1 providers like Cogent and Arelion.

The divergent paths *without* CitMs all follow a similar pattern: once the traceroutes enter the U.S., they stay inside the U.S. until reaching their destinations. The divergent paths *with* CitMs have destinations hosted in two countries: Sweden (Optimizely AB, AS30811) and Ireland (Amazon, AS16509). These paths fall into the following two patterns: (1) the traffic is first routed by the Tier-1 transit provider Arelion into the U.S. and then to Denmark and Sweden, and (2) the traffic is routed by Cogent directly from Canada to the U.K. and finally to Ireland. The divergent paths with CitM only include the U.S. and U.K. before reaching their destination in Ireland or Sweden.