Quantifying Nations' Exposure to Traffic Observation and Selective Tampering

Alexander Gamero-Garrido
CAIDA, UC San Diego

Esteban Carisimo
Northwestern University

Shuai Hao
Old Dominion University

Bradley Huffaker
CAIDA, UC San Diego

Alex C. Snoeren
UC San Diego

Alberto Dainotti
CAIDA, UC San Diego

ABSTRACT

Almost all popular Internet services are hosted in a select set of countries, forcing other nations to rely on international connectivity to access them. We infer instances where traffic towards a large portion of a country is serviced by a small number of Autonomous Systems, and, therefore, may be exposed to observation or selective tampering. We introduce the Country-level Transit Influence (CTI) metric to quantify the significance of a given AS on the international transit service of a particular country. By studying the CTI values for the top ASes in each country, we find that 32 nations have transit ecosystems that render them particularly exposed, with traffic destined to over 40% of their IP addresses privy to a single AS. In the nations where we are able to validate our findings with in-country operators, we obtain 83% accuracy on average. In the countries we examine, CTI reveals two classes of networks that play a particularly prominent role: submarine cable operators and state-owned ASes.

1 INTRODUCTION

Computer networks play a central role in the transmission of information across the world. The goal of this study is to identify instances where a significant fraction of a country's inbound international traffic is managed by a select few networks. Such networks are in a position to observe and tamper with a nation's traffic, as could any third-parties who infiltrate them. For instance, observation (of unencrypted traffic and metadata) may be performed by domestic or foreign actors with the purpose of conducting surveillance or espionage, respectively. Conversely, selective tampering—for instance, with individual network flows carrying popular-application traffic—has been reported by both domestic (e.g., government censorship) and foreign (e.g., dis-information campaigns) actors.

Because actual traffic information is difficult to obtain at a global scale, we instead quantify the fraction of a country's (reachable) IP addresses exposed to tampering and observation by specific networks. While all IP addresses are clearly not created equal, they facilitate an apples-to-apples comparision across nations. Traffic towards any given IP address is handled by so-called transit networks, *i.e.,* those who sell connectivity to the rest of the Internet to other, customer networks for a fee; customers include consumer-serving access networks. Transit networks are frequently invisible and unaccountable to end users. This opacity may allow both domestic and foreign actors to observe or tamper with traffic—capabilities we term *transit influence*—without facing diplomatic or political backlash from governments, activists or consumer groups. We aim to bring transparency to the public regarding oversized observation and tampering capabilities granted to specific transit networks in a large group of nations.

In order to reveal these crucial, nation-level topological features, we develop the country-level transit influence (CTI) metric. CTI quantifies the transit influence a particular network exerts on a nation's traffic. Studying transit influence requires an analysis of the global routing ecosystem which enables networks to exchange traffic between them. The Border Gateway Protocol (BGP) is the central system by which networks exchange interconnection information. CTI is based on an analysis of a large compendia of BGP data [10, 59] and includes both topological and geographic filters aimed at extracting transit influence inferences from incomplete and biased data [30, 35, 51].

CTI is particularly salient in countries that lack peering facilities such as Internet exchange points (IXPs) at which access networks might connect directly with networks of other nations. In these nations, transit networks—often a select few based in geographically distant nations [19, 32, 36, 63]—serve as the dominant form of connectivity to the global Internet. Moreover, the lack of domestic co-location facilities places these nations at further risk of exposure to observation and tampering because popular content is generally hosted abroad [21, 31, 40, 54, 65].

We employ a two-stage approach based on passive inference and active measurement to *(i)* identify transit-dominant (i.e., at-risk) countries, and *(ii)* quantify transit influence of the networks serving them. We validate our findings from both stages with in-country network operators at 123 ASes in 19 countries who confirm that our results are consistent with their understanding of their countries' networks. Our contributions include:

(1) A new Internet cartography metric that quantifies the transit influence a particular network exerts on a nation's traffic: the Country-level Transit Influence (CTI) metric, which ranges over $[0, 1]$.
(2) We apply CTI to infer the most influential transit networks in 75 countries that rely primarily on transit for international connectivity. These countries have, in aggregate, ≈1 billion Internet users (26% of the world [4]).
(3) We find that many of these countries have topologies exposing them to observation or tampering: in the median case, the most influential transit network manages traffic towards 34% of the nation's IP addresses.
(4) We identify two classes of ASes that are frequently influential: those who operate submarine cables and companies owned by national governments.

Ethical disclaimer. We acknowledge several ethical implications of our work. Our mass (validation) survey of operators was classified as exempt by our IRB. Our reporting of available paths to repressive countries might trigger government intervention to remove such paths. Another potential issue is the identification of networks (and specific submarine cables) that would yield the most expansive observation/tampering capabilities in a country, which is potentially useful information for a malicious actor. We believe most governments and sophisticated attackers already have access to this information, and that our study may lead to mitigation of these concerning topological features; thus, the benefits significantly exceed the risk.

Roadmap. The remainder of the paper is organized as follows. We start in Sec. 2 with a high-level overview of our methodology before introducing the CTI metric in Sec. 3. We apply CTI to 75 countries where international connectivity is predominantly transit and describe our findings in Sec. 4. Then, we discuss in detail how we identified the transit-dominant countries (Sec. 5); we also describe how we assign nationality to prefixes, ASes, and BGP vantage points (Sec. 6). Sec. 7 discloses some limitations of our study while Sec. 8 compares with prior work. Space constraints force us to leave a detailed discussion of our extensive operator validation (App. A), quantitative comparision to Hegemony, a closely-related metric [35] (App. B), further details of our methology (App. C–D), and a full discussion of submarine-cable operators (App. E) to the appendices.

2 APPROACH OVERVIEW

Conceptually, international Internet traffic crosses a nation's border at some physical location, likely along a link connecting two routers. For our purposes, we are not interested in the physical topology, but the logical one: in which autonomous system(s) does international traffic enter a nation on its way to access networks in that country (i.e., origin ASes). Topologically, these ASes can have two different types of relationship with the first domestic AS encountered: transit (provider-to-customer or *p2c*) or peering (peer-to-peer or *p2p*). We focus on countries where international connectivity is dominated by transit (p2c) interdomain relationships as they are easier to identify from public data sources.

High-level model. We look for evidence of a country's exposure to observation or selective tampering by specific networks. Studying this exposure requires a quantitative model of the reliance of the country's access networks, in aggregate, on specific transit networks. The model must factor in the size of the address space originated by each AS with presence in the country. Intuitively, the greater the share of a country's IP addresses that are served by a particular transit AS, the higher the potential exposure of the nation's inbound traffic to observation or tampering by that AS. The model must then produce a country-level metric of exposure for each transit network serving the nation. To that end, we determine the frequency at which transit networks appear on routes towards the country's IP addresses.

We start our model by building a graph where nodes are ASes and edges are connections between them, weighted by address space. Then, a metric of node prominence on said graph provides a quantitative assessment of how frequently a (transit) node AS_t is traversed when delivering traffic from any given node to edge (origin) nodes. The higher the value of this metric for any AS_t in a given country, the more exposed the transit ecosystem is. At one extreme (most exposed) are countries with a single transit provider (*e.g.,* a legally-mandated monopoly) connecting every network in the country to the rest of the Internet; at the other end are countries with many transit providers, each delivering traffic to a small fraction of the nation's IPs. Note that we do not need complete visibility of the graph (*e.g.,* backup links) to infer potential exposure to observation or tampering, as traffic will likely flow through the links that are visible given capacity constraints on long-haul (incl. international) links [15, 47, 53, 74].

Our technical approach to build this conceptual model using real data uses as inputs a combination of two types of measurements: *(i)* passive, to study AS-level connectivity, and *(ii)* active, to study reachability and transit dominance.

AS-level connectivity. Passive data sources of interconnection include BGP data from RouteViews [10] and RIPE

RIS [8]. We begin with the 848,242 IPv4 prefixes listed in CAIDA's Prefix-to-Autonomous System mappings derived from RouteViews [24], excluding the 6,861 (0.8%) prefixes with (invalid) length greater than 24, and the 9,275 (1.1%) originated by multiple ASes. We find those prefixes in the 274,520,778 IPv4 AS-level paths observed in BGP table dumps gathered by AS-Rank [11] from RIPE/RouteViews [10][8] during the first five days of March 2020. We consider the set of prefixes and the ASes that originate them on each observed path in combination with the 377,879 inferred AS-level relationships published by CAIDA [7].[1]

Reachability. While BGP dumps reveal potential paths toward destination ASes, they may not reflect the actual routes packets traverse, both because ASes may have alternative routes they do not export (e.g., based on peering relationships) and because the destination network may not, in fact, be reachable. Hence, we conduct a two-week-long active measurement campaign (see Sec. 5.2) in May 2020 to determine which ASes in a country are actually reachable, and the set of ASes traversed by our probe packets (as inferred by BdrmapIT [52]).

In practice, our CTI metric combines BGP data with reachability information for each origin AS included in our large-scale `traceroute` campaign. If at least one of our traceroutes received a response, the origin AS is included in the potentially exposed set of origin ASes. All origin ASes' addresses (including unresponsive networks) in a country are factored in the calculation of the country's total address space. This method yields a conservative estimate of the capabilities for observation and tampering of any given transit AS (an exposure lower bound). In the countries we study using CTI, addresses originated by responsive ASes represent, in aggregate, a median of 88% and an average of 92% of the country's addresses; that figure is over 60% in all but one country (Chad, 56%).

Transit dominance. Because we are focused only on countries where transit—as opposed to peering—is the main form of trans-border connectivity, we use our active campaign to identify and exclude nations with evidence of foreign peering, *i.e.,* where an AS that originates addresses geolocated to the country establishes a peering agreement with another AS primarily based in another country[2].

[1] In the 75 countries where we study transit influence, no path contained any of: unallocated ASes, loops, poisoned paths (where a non-clique AS is present between two clique ASes, clique being the AS-level core of the Internet inferred by [7]); additionally, all paths towards these countries are seen at least once per day across all five days.

[2] This "nationality" assignment is described in Sec. 6.

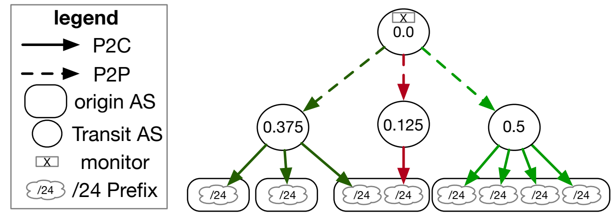


Figure 1: Example of Country-Level Transit Influence.

3 TRANSIT INFLUENCE METRIC

We define the transit influence $CTI_M(AS, C) \in [0, 1]$ using a set of BGP monitors M as

$$\sum_{m \in M} \left(\frac{w(m)}{|M|} \cdot \sum_{p|\text{onpath}(AS,m,p)} \left(\frac{a(p,C)}{A(C)} \cdot \frac{1}{d(AS,m,p)} \right) \right), \quad (1)$$

where $w(m)$ is monitor m's weight (Sec. 3.1.2) among the set of monitors (Sec. 3.2.2); onpath(AS, m, p) is true if AS is present on a preferred path observed by monitor m to a prefix p originated by a probed and responsive origin network, and m is not contained within AS itself (Sec. 3.2.1); $a(p, C)$ is the number of addresses in prefix p geolocated to country C; $A(C)$ is the total number of IP addresses geolocated to country C; and $d(AS, p, m)$ is the number of AS-level hops between AS and prefix p as viewed by monitor m (Sec. 3.1.1).

We provide the intuition behind the derivation of CTI in App. D; here, we illustrate its use. Fig. 1 shows CTI values for a toy example with three transit ASes and four origin ASes, in a country with eight /24 prefixes: the transit AS on the right has the highest CTI, since it serves the most addresses (half of the country), followed by the transit AS on the left (3/8) and the AS in the center (1/8). Note that the top AS has a CTI of 0, because it hosts the BGP monitor from which the set of routes used in this toy example are learned—hence, onpath(AS_t, m, p) is always false for that AS. Should that AS not be the host of the BGP monitor (or be seen on these routes through another monitor), it would have a CTI of 0.5—transit influence over the entire country as an indirect transit provider (distance 2 from the prefixes).

3.1 CTI components

Because Eq. 1 only considers prefixes originated by networks that we probe and are responsive, yet divides by $A(C)$ (*i.e.,* all addresses originated from the country), our experiments yield a conservative estimate. In particular, the actual CTI of AS might be higher if other origin networks that we do not probe are also reached through AS. Moreover, because CTI is computed with respect to the entire country regardless of the amount of probing, it is possible to compare CTIs across countries. Originating addresses directly does not grant an

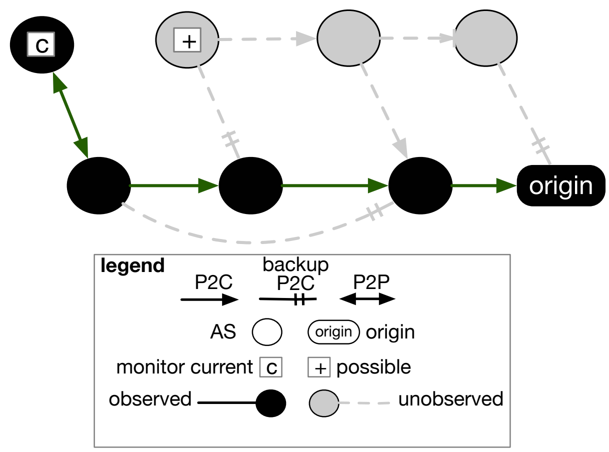


Figure 2: Unobserved paths in BGP.

AS transit influence, as our focus is on identifying ASes that carry traffic to destinations outside of their network.

We explain the rationale for the various factors in Eq. 1 in the following subsections.

3.1.1 Indirect transit discount. As the number of AS-level hops from the origin increases, so too does the likelihood that there exist alternative paths towards the same origin AS of which we have no visibility (*e.g.,* backup links, less-preferred paths). Fig. 2 shows this limitation in visibility for a toy example with a single origin AS. There, given the location of BGP monitor *C* we see the AS-level chain in black, erroneously concluding that the origin AS has a single direct transit provider and two indirect transit providers. In reality, there exists another set of both direct and indirect transit providers (the AS-level chain in light gray). We miss all these paths given that we do not have a monitor in any neighbor of a light-gray AS (such as that marked with a plus sign). In this example we miss backup links of the origin AS, as well as preferred links of the origin's direct transit provider, and a backup link of both indirect transit providers.

As a coarse mechanism aimed at mitigating this limited visibility, we discount the influence of transit providers in proportion to the AS-level distance from the origin: we apply a discount factor as $1/1, 1/2, ..., 1/k$, where k is the number of AS-level hops from the origin AS. In practice, that means we do not discount the measurements of direct transit providers, as there the probability of missing a backup or less-preferred link is low. We note that this heuristic yields a conservative estimate of the observation opportunities of an indirect transit provider over traffic flowing towards a country.

3.1.2 Prioritizing AS diversity. ASes can host more than one BGP monitor. In fact, more than 20 ASes in RIPE RIS and RouteViews host multiple monitors; for instance, AS3257-GTT hosts five. In order to favor a topologically-diverse view (given the available observations), if more than one monitor from the same AS sees an announcement for the same prefix, we discount their observations to limit the influence of monitor ASes with multiple monitors. Formally, the weight for each monitor m's observation of a prefix is $w(m) = 1/n$, where n is the number of BGP monitors in a single AS that see an announcement of that prefix.

3.2 Filtering ASes

To correct for the limited, non-uniform coverage of the BGP monitors that collect our table dumps, we apply a number of filters to the set of paths over which we compute CTI.

3.2.1 Provider-customer AS filter. BGP monitors by definition collect paths from the AS hosting the monitor to the origin AS. Therefore, we always exclude the AS hosting the BGP monitor from the path to avoid inflating their transit influence. We employ a heuristic that attempts to consider only the portion of the path relevant to the origin prefix, and ignore the portion dictated by the monitor's topological location.

The intuition behind our filter is that, from the perspective of the origin AS, there is a "hill" above it capped by the last observed provider-customer (p2c, *i.e.,* transit) link, with traffic flowing from the hill's peak down towards the origin. The transit AS in that link is the highest point in the path we want to keep, as it directs traffic towards its customer (and its customer's customers, if applicable). After reaching that topological peak, we discard any other AS present in the path. The remaining path would then include the origin AS, its direct or indirect transit provider at the topological peak, and any other ASes appearing between the origin AS and the direct or indirect transit provider.

Formally, for the analysis presented in this paper, we refine onpath(AS_t, m, p) to be true only if the path observed at monitor m has at least one inferred p2c link where the customer is either the origin of p or closer to it than AS_t, *i.e.,* we discard paths where there is no topological peak from the perspective of the origin. This heuristic discards 4.3% of the paths observed by our monitors. In the median country we discard 3.4% of paths using this filter, with 6.0% being the average case. In all countries we keep over 78% of paths.

We call this mechanism the *p2c filter*, and it ensures that at least one AS (the inferred customer of the transit AS) relies on at least one other AS (the inferred transit provider) for transit from and towards the core of the Internet. As we aim to measure transit influence, these business relationships are an important source of information: merely being directly connected to an AS path that reaches the origin AS in a given country does not necessarily make an AS influential; being a direct provider of the origin, or of an AS closer to the origin, lends more confidence to our inference of influence.

3.2.2 CTI outlier filtering. We further filter BGP-monitor location noise by removing outlier estimates of transit influence—both overestimates and underestimates resulting from the AS hosting a BGP monitor being topologically too close or too far from the origin AS—to get an accurate assessment of transit influence towards that origin. We implement a filter recently proposed for another AS-topology metric (AS hegemony [35], see Sect. 8). Specifically, we compute the CTI of each transit provider AS_t using BGP monitors from each monitor-hosting AS_h independently, as $CTI_{m(AS_h)}(AS_t, C)$, where $m(AS_h)$ is the set of monitors within AS_h. We determine which potentially-biased AS_h have gathered observations producing $CTI_{m(AS_h)}(AS_t, C)$ values in the bottom and top 10% of all values for that transit provider in that country and disregard all paths observed by monitors hosted in these potentially-biased AS_h. As in [35], we implement outlier filtering only where we have observations of $CTI_{m(AS_h)}(AS_t, C)$ from 10 or more AS_h, which occurs for 52.9% of transit AS-country pairs in our sample (a single AS can operate in multiple countries).

4 COUNTRY-LEVEL TRANSIT

In this section we present the results of applying our CTI metric to the transit ecosystem of 75 countries with little-to-no international peering. (We describe our method for selecting these countries in Sec. 5.) We provide a high-level characterization of the transit ecosystem in each country by comparing the CTI scores of the top-5 ASes ranked by CTI (Sec. 4.1), as well as a set of ASes that appear in the top 5 of many countries (at least 10). Our hypothesis is that these countries show different transit profiles as a consequence of the socioeconomic and geopolitical diversity of the sample: from high exposure to observation, where one AS is the most influential transit provider and others are very marginal, to less exposed countries with an ensemble of ASes with similar values of CTI.

Investigating the companies operating the ASes with high CTI, we find two prominent groups of organizations: submarine cable operators (Sec. 4.2) and state-owned providers (Sec. 4.3). For the former, their operation of physical infrastructure connected to the country lends credence to our inferences.

With regards to state-owned ASes, providing transit may give governments the ability to expand their footprint beyond addresses they originate, *e.g.*, through a state-owned broadband provider. In some cases, state ownership of a transit provider may follow their investment in a submarine cable or landing station, while in others it may reflect the government's intention to enact censorship. We limit our analysis to the discovery of the transit footprint of the state, without delving into the underlying motives.

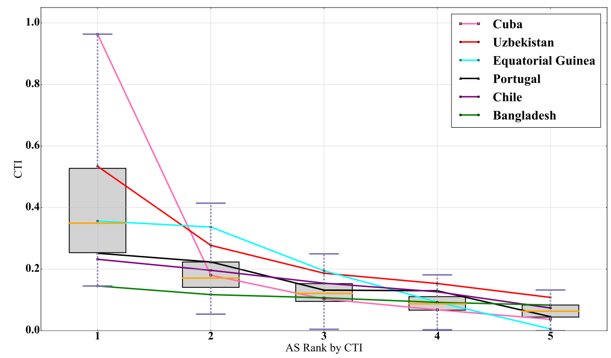


Figure 3: Boxplot of CTI distributions for the top-5 ASes in each country.

4.1 CTI distribution across countries

In this subsection we present an overview of the CTI distribution across countries. Countries with a top-heavy distribution of CTI values are particularly exposed to specific networks. Other nations with a more flat distribution signal an ecosystem that is less exposed to prominent transit ASes. Fig. 3 shows the distribution of CTI values for ASes ranked in the top 5 by CTI in each country. In 51 countries, the top-ranked AS has CTI ≥ 0.3, signaling high exposure to observation and tampering by that specific network.

The distribution of CTI rapidly declines across AS rank, with the median halving from the first to the second position. In 54/75 countries, CTI declines by over 30% from the top-ranked AS to its successor; the average and median decline across all countries are 46% and 49%. This suggests that in the vast majority of countries in our sample, a single AS is particularly prominent in terms of its capabilities to observe or tamper with traffic.

4.1.1 Individual nations. Results for the full set of countries we study are included in Table 1. In the interest of parsimony, we discuss several representative cases below.

Most exposed countries. Only five countries have a top-ranked AS with a CTI over 0.75: Cuba, Libya, Sierra Leone, Solomon Islands and Cape Verde. The latter two are small island nations. Among the remaining countries, Cuba appears to have the most-exposed transit ecosystem[3], in which the top-ranked AS has CTI of 0.96. Because CTI discounts indirect transit—and the top AS monopolizes observed, direct connectivity—the CTI of Cuba's remaining ASes declines rapidly (81% from the top-ranked AS to the second).

[3]This is consistent with previous work that focused exclusively on Cuba, finding its international connectivity to be constrained [19].

Country name	CC	AS1	AS2	AS3	AS4	AS5	HHI
CAPE VERDE	CV	0.76	0.12	<0.01	<0.01	-	0.69
MALI	ML	0.7	0.13	0.07	0.03	0.02	0.46
SRI LANKA	LK	0.53	0.05	0.05	0.05	0.05	0.45
CUBA	CU	0.96	0.18	0.1	0.07	0.04	0.42
ST VINCENT	VC	0.67	0.11	0.1	0.07	0.01	0.41
LIBYA	LY	0.95	0.41	0.05	0.01	0.01	0.4
SAINT LUCIA	LC	0.59	0.1	0.1	0.08	0.02	0.34
SOLOMON ISLANDS	SB	0.78	0.39	0.05	0.03	0.02	0.34
BARBADOS	BB	0.59	0.12	0.11	0.06	0.05	0.31
MOROCCO	MA	0.59	0.14	0.07	0.07	0.06	0.31
ZAMBIA	ZM	0.58	0.15	0.07	0.07	0.04	0.3
IRAQ	IQ	0.57	0.14	0.09	0.08	0.05	0.28
GUYANA	GY	0.54	0.3	0.06	0.04	0.03	0.27
SIERRA LEONE	SL	0.81	0.37	0.11	0.08	0.06	0.25
BELIZE	BZ	0.34	0.07	0.07	0.06	0.03	0.24
YEMEN	YE	0.48	0.31	0.08	0.05	0.04	0.21
EL SALVADOR	SV	0.46	0.18	0.11	0.05	0.04	0.21
TURKMENISTAN	TM	0.33	0.24	0.07	0.03	0.02	0.21
TRINIDAD AND TOBAGO	TT	0.56	0.15	0.11	0.09	0.09	0.21
OMAN	OM	0.51	0.15	0.13	0.1	0.04	0.21
BOLIVIA	BO	0.54	0.27	0.11	0.08	0.05	0.19
PERU	PE	0.42	0.14	0.1	0.06	0.05	0.19
JORDAN	JO	0.55	0.15	0.13	0.1	0.09	0.19
LUXEMBOURG	LU	0.3	0.11	0.08	0.06	0.04	0.16
NAURU	NR	0.55	0.28	0.17	0.09	0.05	0.16
TUVALU	TV	0.6	0.3	0.2	0.13	0.04	0.14
JAMAICA	JM	0.51	0.17	0.12	0.12	0.11	0.14
ST KITTS	KN	0.33	0.12	0.07	0.07	0.07	0.14
PANAMA	PA	0.44	0.19	0.14	0.08	0.07	0.13
LESOTHO	LS	0.48	0.2	0.19	0.11	0.06	0.13
EGYPT	EG	0.37	0.14	0.1	0.1	0.07	0.12
ETHIOPIA	ET	0.58	0.26	0.25	0.1	0.08	0.12
ESWATINI	SZ	0.25	0.08	0.07	0.07	0.06	0.12
EQUATORIAL GUINEA	GQ	0.34	0.32	0.19	0.09	0.01	0.12
NICARAGUA	NI	0.31	0.14	0.09	0.07	0.06	0.11
MONGOLIA	MN	0.46	0.16	0.14	0.11	0.11	0.11
SUDAN	SD	0.46	0.15	0.14	0.13	0.1	0.11
MONTENEGRO	ME	0.44	0.18	0.14	0.11	0.09	0.11
ZIMBABWE	ZW	0.38	0.21	0.13	0.08	0.07	0.11
MYANMAR	MM	0.3	0.13	0.13	0.07	0.05	0.11
CAMEROON	CM	0.44	0.17	0.17	0.11	0.09	0.11
DRC	CD	0.27	0.15	0.09	0.06	0.06	0.1
BAHAMAS	BS	0.32	0.2	0.17	0.07	0.04	0.1
QATAR	QA	0.27	0.19	0.09	0.07	0.06	0.09
BURKINA FASO	BF	0.32	0.28	0.22	0.05	0.05	0.09
UZBEKISTAN	UZ	0.46	0.24	0.16	0.13	0.09	0.09
ARMENIA	AM	0.33	0.18	0.12	0.09	0.07	0.09
KUWAIT	KW	0.32	0.14	0.14	0.13	0.05	0.08
HAITI	HT	0.37	0.17	0.16	0.1	0.09	0.08
MALTA	MT	0.31	0.18	0.18	0.08	0.06	0.07
KOREA	KR	0.21	0.15	0.08	0.07	0.06	0.07
GUATEMALA	GT	0.22	0.22	0.17	0.07	0.04	0.07
THAILAND	TH	0.34	0.19	0.17	0.14	0.06	0.06
GEORGIA	GE	0.25	0.16	0.11	0.08	0.07	0.06
PALESTINE, STATE OF	PS	0.22	0.18	0.14	0.1	0.03	0.06
COLOMBIA	CO	0.33	0.18	0.15	0.11	0.1	0.06
INDIA	IN	0.25	0.18	0.17	0.12	0.03	0.06
SAN MARINO	SM	0.27	0.14	0.12	0.09	0.09	0.06
PORTUGAL	PT	0.26	0.23	0.13	0.13	0.05	0.06
SOMALIA	SO	0.26	0.17	0.14	0.1	0.08	0.05
TIMOR-LESTE	TL	0.27	0.27	0.23	0.13	0.07	0.04
TONGA	TO	0.35	0.24	0.17	0.15	0.13	0.04
GUINEA	GN	0.23	0.21	0.21	0.1	0.07	0.04
AFGHANISTAN	AF	0.22	0.15	0.12	0.1	0.07	0.04
VENEZUELA	VE	0.31	0.19	0.15	0.15	0.13	0.03
CHILE	CL	0.26	0.22	0.17	0.14	0.08	0.03
HONDURAS	HN	0.21	0.16	0.1	0.1	0.09	0.03
ECUADOR	EC	0.2	0.19	0.13	0.12	0.07	0.03
ALBANIA	AL	0.22	0.16	0.13	0.12	0.09	0.03
SAMOA	WS	0.27	0.23	0.19	0.18	0.09	0.02
NORTH MACEDONIA	MK	0.19	0.19	0.13	0.11	0.09	0.02
TURKEY	TR	0.16	0.13	0.12	0.09	0.08	0.01
CHAD	TD	0.14	0.12	0.11	0.09	0.08	0.01
BANGLADESH	BD	0.15	0.12	0.11	0.09	0.08	0.01
BELARUS	BY	0.15	0.13	0.11	0.11	0.09	0.01

Table 1: CTI values of the five top-ranked ASes in the 75 countries we study (colored rows match Fig. 3). Table is sorted by the Herfindal-Hichman Index (HHI) [3] of the (normalized) top-5 CTI values. This index reflects both the exposure of the country to the top ASes, in aggregate, as well as the skew of the exposure's distribution.

Countries around the median. The median of the leftmost bar in Fig. 3 consists of countries that are still considerably exposed to observation and tampering, with CTI values ranging from 0.34 to 0.37: Egypt, Tonga, Equatorial Guinea, Belize and Thailand. In Eq. Guinea, the median country, the top-two ASes each have a CTI over 0.3; these ASes have a p2c relationship with each other. Egypt and Belize have more skewed distributions, with a 62–79% decline from the top AS to its successor.

Least exposed countries. At the other end of the spectrum in Fig. 3 are five countries where the top-ranked has CTI values under 0.2: Chad, Bangladesh, Belarus, Turkey and North Macedonia. These countries have flatter distributions, with CTI declining at most 19% (or 13% on average) between the top-two ASes. As a result, we find no evidence of these nations being particularly exposed to a single network (unlike most of their peer countries in our sample). India, the country with the most Internet users in our sample, is in the bottom quintile (close by the other nations mentioned in this paragraph) with a top-AS CTI of 0.25, declining by 28% between the top 2 ASes.

Frequently Top-Ranked ASes. Of the 170 ASes present in Fig. 3, 129 of them are in the top-5 for only one country, with a further 34 ASes in the top-5 of at most 10 countries. There are some notable exceptions, however: 1299*-Telia (top-5 in 26 countries), 174*-Cogent (25), 3356*-Lumen (formerly Level3/CenturyLink) (22), 6939-HE (17), 6762*-T. Italia (14), 23520-C&W (14), and 6453*-Tata (12). Nearly all of these networks (marked with *) are in the inferred clique at the top of the global transit hierarchy [1]. C&W is only present in our analysis for countries in the Caribbean. HE has a very broad footprint, with countries in Africa (7), the Mid. East (3), W. Europe (2), Southeast Asia (2), South Pacific (2) and E. Asia (1).

4.1.2 Temporal stability. We apply our CTI methodology to a set of BGP paths from Feb. 2020 and compare the results to those discussed above (from Mar. 2020). Specifically, we compute the absolute value of the difference in CTI across both months for ASes listed in the top 5 for each country in Mar. 2020. We compute the absolute difference in CTI for a total of 374 AS-country pairs[4], or 172 ASes in 75 countries. The 25th ptile., mean, median, and 75th ptile. of this absolute difference are 0.002, 0.003, 0.008 and 0.025, so the CTI values are relatively stable across these months.

4.2 Submarine cable operators

Submarine cables are known to be an important part of the global Internet infrastructure [18, 33, 48] and play a role in the top-5 ASes of most countries we study. (Nicaragua,

[4] Cape Verde only has four transit ASes, which is why there are 374 AS-country pairs instead of 375.

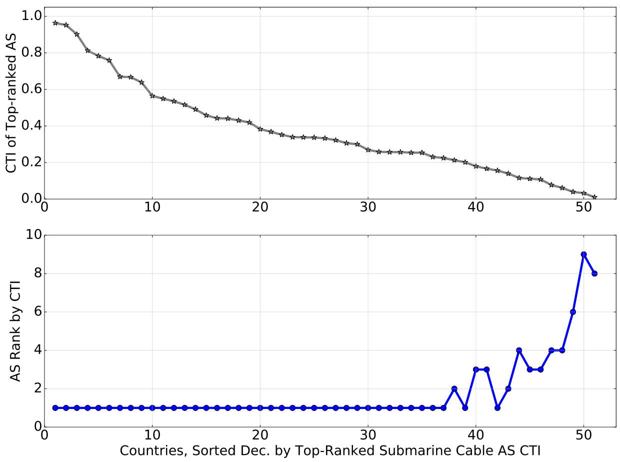


Figure 4: Top: CTI of top-ranked submarine cable AS. Bottom: CTI rank of top-ranked submarine cable AS.

Guatemala, and Guyana are the only three nations where none of the top-5 ASes are associated with the submarine cables landing in the country.) In this section, for each country, we find the highest-ranked AS by CTI where there is evidence of an institutional connection between the AS and an owner or operator of a submarine cable. We define an AS as a submarine cable operator if we find a direct match between the AS Name, the AS Organization [23], or a corporate parent organization (*e.g.,* CenturyLink for Level3, the Government of Sierra Leone for Sierra Leone Cable Company) and the owners of a submarine cable operator according to TeleGeography [69] and Infrapedia [41].

This process yields submarine cable ASes in 46 countries out of 51 possible, as 19 of the 75 countries are landlocked, and 5 have no submarine cable connectivity according to the operator databases. In three additional countries (Myanmar [67], Solomon Islands [28], and Congo DRC [46]) only TeleGeography provides an AS to submarine cable match, which we confirm with information from the cited sources (the operators themselves, the government of Australia, and a submarine cable news source). In the remaining two countries (Thailand [71] and Samoa [68]) where we were not able to find an AS to submarine cable from TeleGeography, we rely on the cited sources (from the operator and a Samoan news outlet) to find a match. Note that only operators of submarine cables who appear as an AS on the BGP path can be identified using this method, so our findings may be a lower bound of the influence of submarine cable operators in some countries, *i.e.,* where we do not identify a submarine cable AS ranked first by CTI.

Our findings are shown in Fig. 4, with the CTI of the top cable-owning AS in each of the 51 countries shown in the upper portion, and the ordinal ranking of that AS in its country's ecosystem in the bottom portion (the order of countries is the same in both plots, and sorted by the CTI of the top cable-owning AS). In 39 countries, a submarine cable AS is ranked at the top by CTI, with an average rank of 1.8.

Note that being the top operator by CTI means different things in different countries, as the underlying potential exposure to observation affects the CTI of the top AS. For instance, in Turkey and South Korea a cable-owning AS ranks first by CTI, but has the lowest CTI among such countries. Said ASes (9121-Turk Telecom and 6939-Hurricane Electric) have CTIs of 0.16 and 0.21, respectively. By contrast, in Cuba and Libya, a submarine cable operator (11960-ETECSA and 37558-LIT) is also ranked first but with CTIs of 0.96 and 0.95, respectively. As a result, Turkey and South Korea are much less exposed to a single AS than Cuba and Libya.

We also find regional clusters of high transit influence for the same AS operating a submarine cable, including C&W (formerly Columbus Networks), which is among the top providers in 10 countries in Central America and the Caribbean thanks to its ownership of the ECFS and ARCOS-1 cables. Tata, Telefonica and Bharti Airtel also have an important transit presence in West Africa, Western South America, and South Asia respectively. A complete list of submarine cables linked to an AS with high CTI in multiple countries is included for reference in App. E.

4.3 State-owned transit providers

In more than a third (26) of nations, we find that at least one of the top-5 ASes is state-owned (according to a recent study [16]), motivating us to further examine the total influence of a country's government on its Internet connectivity. In particular, we adapt CTI to quantify the influence of state-owned conglomerates—as some nations have more than one state-owned AS—and apply it to the 75 countries in our sample. We use as input a list of ASes that are majority-owned by sovereign states [16]. The list was manually verified and encompasses both access and transit ASes. The dataset includes major telecommunication providers as well as its sibling networks and subsidiaries. Using this list, we find 100 state-owned ASes who operate domestically (*i.e.,* where the state owner and the country of operation are the same) in 41 countries.

4.3.1 Influence of State-Owned ASes. Our initial exploration of the influence of state-owned ASes concerns the role each AS plays in the ecosystem of its country, as shown in Fig. 5. We find that state-owned ASes tend to provide either transit or access, usually not a combination of both. (Most points in Fig. 5 line up along an axis, rather than towards the middle.) As a consequence, meaningfully estimating the footprint of the state requires combining the two kinds of influence as well as aggregating data for AS conglomerates. Two exceptions where a state-owned AS provides both Internet access (*i.e.,* as an origin AS) and serves transit to other ASes are Cameroon and Egypt; in the former, Camtel has both a high CTI (0.44, ranked first) and originates 27% of the country's addresses (second only to Orange Cameroon). Egypt's TE has a CTI of 0.37 and originates 28% of the country's addresses.

We begin that estimation by computing CTI for not just a single AS, but a set of ASes, while not "double counting" influence over the same addresses, *i.e.,* if two of the state's ASes originate and provide transit to the same addresses, we add those addresses to the state's footprint once. We call this derived metric $CTIn$. Intuitively, $CTIn$ reflects the "pure-transit" footprint of the state, crediting only the addresses where state-owned ASes serve exclusively as transit providers. For instance, if AS A and AS B (both of which operate in country C) respectively originate and provide transit to the same /24 prefix, $CTIn$ says that the conglomerate $S_C = \{A, B\}$ does not have transit influence over the /24 prefix. Formally, $CTIn_M(S_c, C) \in [0, 1]$ is calculated as

$$\sum_{m \in M} \left(\frac{w(m)}{|M|} \cdot \sum_{p|\text{onpath}^*(S_c,m,p)} \left(\frac{a(p,C)}{A(C)} \cdot \frac{1}{d^*(S_c, m, p)} \right) \right),$$

which is essentially identical to Eq. 1, except that S_c is a set containing all of the ASes in the state-owned conglomerate of country C; onpath$^*(S_c, m, p)$ is true if onpath(AS_t, m, p) is true for some $AS_t \in S_c$ and p is *not* originated by any AS in S_c; and $d^*(S_c, m, p) = \min_{AS_t \in S_c} d(AS_t, m, p)$, *i.e.,* the AS-level distance from p to the closest AS in the conglomerate.

Finally, we define the total footprint of the state, *i.e.,* addresses that are either originated or for which transit is served by a state-owned AS. The state's footprint $F(C) \in [0, 1]$ is calculated as

$$F(C) = CTIn_M(S_c, C) + \sum_{AS_o \in S_c} \frac{a^*(AS_o, C)}{A(C)},$$

where $a^*(AS_o, C)/A(C)$ is the fraction of addresses in country C originated by AS_o. The first term of the sum is the pure-transit footprint and the second term is the addresses directly originated by the state-owned conglomerate S_c.

4.3.2 Findings. Fig. 6 shows our findings for the state-owned footprint (F, bar height), the originated fraction by state-owned ASes (orange bar), and pure-transit footprint of state-owned ASes ($CTIn$, blue bar).

Our results suggest that domestic state influence exists on a spectrum where some countries, such as Ethiopia, Cuba, Libya and Yemen, rely overwhelmingly on the state for the provision of Internet access and (F between 0.90–0.97), whereas others, such as Colombia, Turkey, Mongolia and Ecuador have relatively marginal state-owned enterprises (F between 0.01–0.12).

Regarding the mode of influence that states use, in many countries in Fig. 6, most of the bar height is contributed by

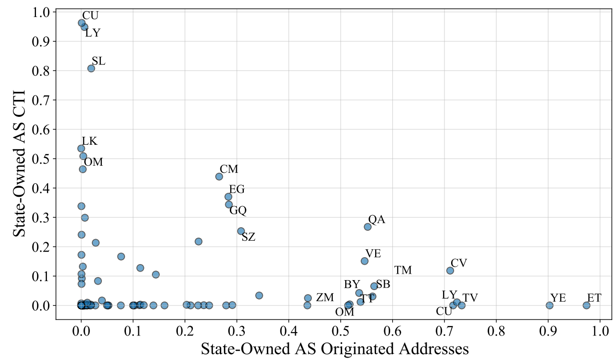


Figure 5: CTI and fraction of addresses originated by domestic, state-owned ASes in our study.

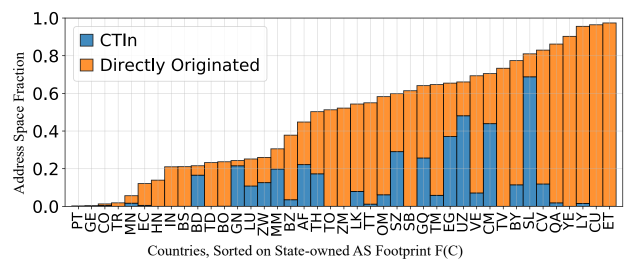


Figure 6: State-owned originated address space a^* (orange bars), $CTIn$ (blue bars), and state footprint F (bar height) for countries in our study (X-Axis, sorted by F).

the orange portion, meaning that the footprint of the state comes from addresses directly originated. However, in some countries the state punches above its access network weight by deploying an influential transit provider, *i.e.,* those where the bar height is not dominated by the origin contribution in orange.

The countries where pure-transit influence ($CTIn$) is largest (0.2 or more, or pure-transit influence over at least a fifth of the country's addresses) are shown in Tab. 2. In these countries, all of which are in Africa and Central Asia, providing transit considerably increases the influence of the state.

4.3.3 Pure-transit footprint of state-owned ASes. In countries where state-owned ASes have a large value of $CTIn$ (Tab. 2), it is possible that providing Internet access directly is beyond the capabilities of the state (at least in some of each country's regions) which would explain the relatively low footprint contribution of addresses directly originated. In these countries, building an influential transit network may be a cost effective way to expand the purview of the state, be it for monetary gain (improved tax collection), infrastructure

Country	$CTIn$	F
Sierra Leone	0.69	0.81
Uzbekistan	0.48	0.66
Cameroon	0.44	0.71
Egypt	0.37	0.65
Swaziland	0.29	0.60
Eq. Guinea	0.26	0.64
Afghanistan	0.22	0.45
Guinea	0.22	0.24
Myanmar	0.20	0.31

Table 2: Top countries by $CTIn$.

improvement (increasing the country's available international bandwith), or surveillance (expanding the fraction of the country's traffic that traverses a state-owned organization). We note that the mere existence of these influential transit ASes does not signal willingness of the state to engage in surveillance or selective tampering, but rather that the government may have opportunities to do so. For instance, Myanmar's state-owned *Myanma Posts and Telecommunications (MPT)*, which is included in our analysis (see Tab. 2), appears to have been involved in the disruption of the country's Internet service during the recent coup [39].

4.4 Impact of public policy

In our CTI results, we find anecdotal evidence of the impact of policy on each country's telecommunications ecosystem in two ways. First, while the underlying motives for centralized and state-owned operation of national networks is outside the scope of this study, it is worth noting that the four countries where $F > 0.9$ (Ethiopia, Cuba, Lybia and Yemen) are all labeled as authoritarian countries by the Democracy Index [5], so the national government's extensive footprint may allow for effective surveillance or censorship capabilities. Second, two countries where public policy has generally favored the diversification of international routes (Bangladesh [55]), and the establishment of a strong domestic peering mesh (Chile [17, 26, 56]), have among the lowest values for CTI of the top-ranked AS (0.15 and 0.26, respectively, or the bottom quartile of the countries in our sample). These trends, if emulated in other nations, might mitigate the risks imposed by concentration of inbound routes on a few ASes.

5 INFERRING TRANSIT DOMINANCE

In this section, we describe how we identified the 75 countries that where the focus of the preceding section, i.e., countries where provider-customer transit (p2c) relationships are likely the dominant mode of inbound international connectivity.

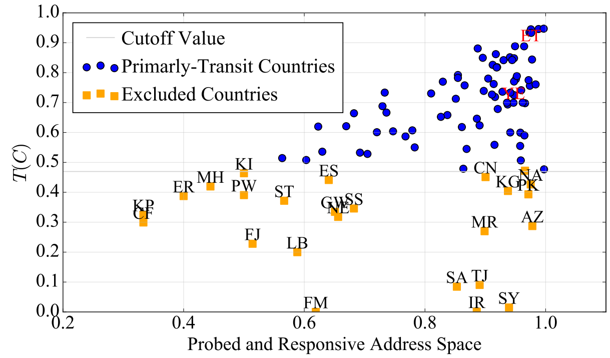


Figure 7: Country-level transit fractions $T(C)$ for countries in our sample.

We start by identifying countries for which public datasets of Internet Exchange Points (IXPs) and Private Colocation facilities (Colo) show no evidence of international peering (Sec. 5.1). Based on this analysis, we conduct an active measurement campaign to confirm the absence of international peering (Sec. 5.2). This second stage based on traceroutes is necessary because peering datasets are incomplete, particularly when it comes to membership lists at IXPs in developing countries [50]. We consider the prevalence of transit links being used to reach each of our target countries from probes distributed worldwide (Sec. 5.3) to select a set of countries where it is likely valid to apply our transit influence metric.

We define international peering as a (logical) link between two ASes that: *(i)* operate primarily in different countries (Sec. 6), and *(ii)* where that link is not an inferred transit-customer link. We use this definition since we are interested in studying the AS-level routes taken towards each country.

We are aware of the limitations of our measurements and analysis, particularly with regards to the location (both topologically and geographically) of our probes. We expand on this discussion in Sec. 7.

5.1 Constructing a candidate list

We identify countries where international peering may not be prevalent by evaluating evidence of international peering involving origin ASes present in the country. While domestic peering is very common, our hypothesis is that international peering is still not a frequent occurrence in some countries. We begin with the set of ASes that originate at least 0.05% of addresses in each country. (We remove marginal ASes that originate a very small fraction of the country's address space to reduce the scope of our active campaign, as we are limited by RIPE Atlas's system-wide limits on concurrent measurements [60]). This set includes origin ASes that we classified as foreign to that country, but that originate BGP prefixes entirely geolocated in the country. We look for these origin ASes in CAIDA's IXP dataset[5] (from October 2019 [22]), PeeringDB Colo dataset (from March 1st, 2020 [25]), and inferred AS-Relationships from BGP (March 2020 [7]).

We classify an origin AS as a *candidate* if the following three conditions are true:

(1) the origin AS has no foreign peers in BGP [7];
(2) the origin AS is not a member of any IXPs or Colos based in another country [22, 25]; and
(3) the origin AS is not a member of any IXPs or Colos where any member AS is based in a different country than the origin AS [22, 25].

The intuition for each test is as follows. If we observe at least one foreign peer on BGP (1), this origin AS already has the ability to receive some external content from that peer, bypassing transit providers. Therefore, transit providers serving that origin will have fewer capabilities to observe traffic flowing towards it. Further, if an AS is a member of an IXP/Colo in another country (2), or a member of an IXP/Colo where another member is from a different country (3), the origin AS is at least capable of establishing peering relationships with those other ASes.

Fig. 8a shows the percentage of a country's address space originated by candidate ASes. We select the top-100 countries as candidates for active measurements. This set includes only countries where at least 25% of addresses are originated by candidate ASes. Our motivation is to actively probe the set of countries where it is most likely that transit providers still play an important role on inbound international connectivity. These 100 countries are colored in Fig. 8b.

5.2 Active measurement campaign

We ran a traceroute campaign to the 100 candidate countries for 14 days starting May 2nd, 2020. Additionally, we use all publicly available IPv4 traceroutes on RIPE Atlas during the same period—on the order of several million per hour—in order to opportunistically take advantage of other measurements towards the same ASes.

We design our traceroute campaign guided by two constraints. First, we want to select a geographically and topologically diverse set of probes. Second, we have to operate within the rate limits of RIPE Atlas[6], particularly regarding concurrent measurements and credit expenditure.

Within these constraints, we launch ICMP traceroutes[7] from 100 active—shown as "connected" during the previous day [61]—RIPE Atlas probes (located outside any target country) towards a single destination in each AS, twice daily[8]; probing at this frequency gives us 28 opportunities to reach the AS during the two-week period from each vantage point.

We target an IP in a single /24 block for each origin AS in each candidate country by looking for any prefix originated by that AS that is entirely geolocated or delegated within the candidate country (see Sec. 6). Our final dataset is comprised of 33,045,982 traceroutes, including those launched by other RIPE users that meet our constraints. The distribution of the number of traceroutes reaching each country has the following properties: (Min, 25th Pctl., Median, Mean, 75th Pctl., Max) = (36, 13k, 46k, 330k, 250k, 3.3m). That is, the median country received 46k traceroutes. Only three countries received fewer than a thousand traceroutes: Eritrea (667), Nauru (154), and Tuvalu (36).

We use BdrmapIT [52] to translate our traceroutes into AS-level interconnections. BdrmapIT requires a number of external datasets in its operation, which we specify as follows: inferred AS-Level customer cone [51] from March 2020; *AS2Org*, which infers groups of ASes who belong to the same organization[9], from January 2020; and datasets we mention in other sections—prefix-to-Autonomous System mappings (Sec. 2), *PeeringDB* records (Sec. 5.1), and RIR delegation records (Sec. 6). From these traceroutes and external datasets, *BdrmapIT* infers a set of AS-level interconnections and the IP addresses (interfaces) at which they occur. Each interface inferred by *BdrmapIT* has an AS "owner" assignment. We reconstruct the AS-level path observed on the traceroute using such assignments.

5.3 Country-level transit fraction

From the preceding sections we have built a set of AS-level paths taken from the traceroute source to the destination AS. We now need a quantitative analysis technique to infer the prevalence of transit links on inbound traces towards each country.

To that end, we determine how frequently a transit (p2c) link is traversed when crossing the AS-level national boundary[10] towards an origin AS (AS_o) in a candidate country. We infer the AS-level national boundary as the link between the last foreign AS observed on the AS-level path (starting from the vantage point) and the subsequent AS.

We calculate how frequently, in the inbound traceroutes we process with *BdrmapIT*, the AS-level national border crossing occurs on a transit link for each origin AS. We scale this fraction to take into account the size of the address

[5] "This dataset provides information about Internet eXchange Points (IXPs) and their geographic locations, facilities, prefixes, and member ASes. It is derived by combining information from PeeringDB, Hurricane Electric, Packet Clearing House (PCH), and GeoNames." [22]

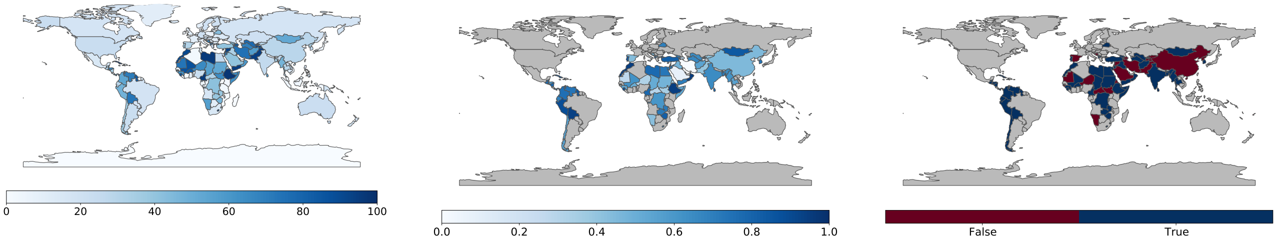
[6] Which RIPE Atlas generously relaxed for this study upon direct request.

[7] Using all default RIPE Atlas values save for number of packets to send, which we reduce to 1 to stay within our measurement credit budget.

[8] Because of limits on user-defined measurements [60] we space traceroutes an hour apart in 800-target IP blocks, which also allows time for the RIPE Atlas server to process and execute each request.

[9] This dataset is not published monthly.

[10] As defined by our AS Nationality, not actual political borders.



(a) (b) (c)

Figure 8: Non-peering observed perc. on passive datasets 8a, scaled country-level transit fraction in probed countries 8b, and final set, with countries in red excluded 8c.

space originated by each AS using the *country-level transit fraction*:

$$T(C) = \sum_{AS_o, AS_c \in \mathrm{dom}(C)} \sum_{AS_t \notin \mathrm{dom(C)}} \frac{R(AS_o, AS_t, AS_c)}{R(AS_o)} \cdot \frac{a^*(AS_o, C)}{A(C)},$$

where $R(AS_o, AS_t, AS_c)$ is the number of traceroutes destined toward a prefix originated by AS_o that traverse a transit link between a foreign provider AS_t and a domestic customer AS_c in country C; $R(AS_o)$ is the total number of traceroutes where AS_o is the last observed AS; and $a^*(AS_o, C)/A(C)$ is the fraction of country C's address space originated by AS_o. For instance, if an AS originates 50% of the country's origin addresses, and 50% of the traces towards it traverse a foreign transit provider AS, the contribution of that AS to the country-level transit fraction becomes 0.25. Note that AS_c and AS_o are not necessarily the same, as the border crossing may occur at the link between (direct and/or indirect) providers of AS_o.

The values of $T(C)$ for each candidate country are represented in Fig. 8b: countries in darker shades of blue have both a large probed and responsive fraction and a large fraction of traceroutes from outside the country traversing transit providers. Countries where this fraction is closer to 1 will be adequately captured by our CTI model of transit influence.

5.4 Final selection

Finally, in order to identify a set of primarily-transit countries, we evaluate the values of $T(C)$ across countries, shown in Fig. 7. At one extreme of Fig. 7 and Fig. 8b are countries such as Ethiopia (ET) and Yemen (YE), $T(C) = 0.95$ and 0.7, respectively, where all available evidence points towards transit links as the main inbound modality. At the other extreme are countries such as Syria (SY) and Iran (IR), $T(C) \leq 0.01$, where we rarely observe AS-level national borders being crossed using transit links.

Outside the upper and lower extremes in Fig. 7, where the decision of whether to include a country in our study is obvious, the middle results (most countries) do not offer clear dividing points. We decided then to set the threshold for $T(C)$ to classify a country as primarily-transit based on our validation with operators (Sec. A); in particular, we use the value of $T(C)$ for Sudan (0.48) as a lower bound, which is the lowest $T(C)$ in any country that we were able to confirm relies on transit links for its inbound connectivity.

The final countries in our CTI study are shown in a blue-white spectrum in Fig. 8c and as blue circles in Fig. 7, 75 of the 100 candidates[11]. Countries in red are excluded from further analysis, as at this time we lack sufficient evidence to support that they are primarily using transit providers for inbound connectivity.

6 DEFINITIONS OF NATIONALITY

Our study hinges on the correct nationality assignment for IP address prefixes, ASes and BGP monitors. Given the diverse set of information available, we devise distinct methods for each. For our purposes, a country is one of the 193 United Nations member states, either of its two permanent non-member observer states, or Antarctica.

Address prefixes. We first geolocate each IP address in every observed BGP prefix to a country using Netacuity [13]. (While geolocation databases are known to be unreliable at fine granularities, previous work has found them to be more accurate at the country level [20, 57], with Netacuity in particular having accuracy between 74–98% [38].) Then, on a country-by-country basis, we count how many addresses in each prefix are geolocated to that country. If the number is less than 256 (a /24), we round up to 256. If Netacuity does not place any of a prefix's IP addresses in a country, we attempt to find a delegation block from the March 2020 RIR delegation files [9] that covers the entirety of the prefix. If there is one we assign all of the delegated prefix's addresses to the indicated country. Hence, while Netacuity can place a prefix in multiple

[11] These geographically small nation-states are included (2-letter ISO code): MT, BS, CV, TO, LU, SB, SM, TV, PS, LC, WS, NR, VC, and KN.

countries, at most one country will receive addresses through the RIR process, and only if it was not already associated with the prefix through Netacuity. Netacuity accounts for 95.1% of all prefix-to-country mappings, while delegation-derived geolocation accounts for the rest.

A particularly pressing concern with geolocation is the correct assignment of IP addresses belonging to large transit ASes with a presence in many countries. We compute the fraction of a country's address space that is originated by ASes that have at least two thirds of their addresses in that country. In the vast majority of countries, the address space is dominated by ASes that are primarily domestic.

BGP monitors. As our study is focused on measuring inbound country-level connectivity, we seek to limit our analysis to paths going towards addresses in the target country from a BGP monitor located outside that country. Hence, we confirm the BGP monitor locations listed by RouteViews [64] and RIPE RIS [62] through a set of active measurements. The details of this process are included in Appendix C.

Autonomous Systems. Our inference of transit-dominant countries relies on a concept of AS nationality, which is rooted in the intuition that an AS will use its IP addresses in countries where it operates, and that the country with most of its addresses will therefore be the primary country of operation. For transit providers, we include the IP addresses originated by direct customers, as they are part of their transit footprint. We exclude indirect customers (*e.g.,* customers of customers) as these do not have a direct relationship with the transit provider, and as a consequence it is possible the two ASes have a peering relationship we do not observe (in which case the indirect customers' addresses would not be part of the transit provider's transit footprint).

We classify each autonomous system AS operating in a country C as being *domestic*, $AS \in \mathrm{dom}(C)$, when the AS has at least two thirds of its addresses in the country, and *foreign* otherwise. The vast majority (97.4%) of ASes are classified as domestic in one country, with the remaining small fraction being classified as foreign in every country. In fact, 89.8% of ASes have all of their address in a single country, and 98.6% have a strict majority of addresses in one country.

7 LIMITATIONS

At a high level, CTI assumes all ASes and IP addresses are equivalent, which is certainly not the case. At the AS level, it is possible that one, dominant AS provides stronger security than a multitude of smaller ASes with tighter budgets. From the perspective of an attacker, though, a single AS having high CTI creates an opportunity; in the case of sophisticated attackers such as nation-states, the possibility of infiltration of any network cannot be discarded, but compromising many ASes simultaneously (in order to observe traffic towards countries where no AS has high CTI) may be more challenging. As such, ASes with very high CTI still present a concerningly large observation footprint, regardless of their level of security against infiltration.

Similarly, IP addresses can represent vastly different entities. Both access and transit ASes may deploy carrier-grade network access translation (CGNAT) [58]. Since our model treats all routed IPs equally, it does not currently take into account the number of hosts multiplexing a single IP address. We leave this to future work, but note that an additional weight may be added to CTI: one that scales up the number of IP addresses in a given prefix by the number of hosts connected to those IPs, on aggregate. Even within a given network, however, individual hosts are unlikely to be equally important as some (e.g., those belonging to governmental organizations or power-grid operators) may have more sensitive traffic. Conversely, some networks might not even actually use all their IP addresses—although the latter issue is likely less of a concern in the countries we have studied as their allocation of IPv4 addresses tends to be constrained [29].

In addition to this fundamental conceptual limitation, there are a variety of technical details that could have out-sized impact on our conclusions:

Incomplete BGP data. We acknowledge that the BGP paths we observe and use to compute CTI are incomplete given the location of BGP monitors. Given the serious implications for countries that appear highly exposed to external observation and selective tampering by an AS, we argue that it is important to study such exposure with available data. Further, we note that there are two important factors aiding the credibility of our CTI findings: *(i)* our validation with network operators, who have confirmed that the set of transit ASes identified in their countries is largely consistent with their own understanding of the country's routing ecosystem. *(ii)* There is greater visibility over p2c links in the AS-level topology [30, 51], which enables our analysis as we are studying exposure to observation or selective tampering by transit ASes, in particular.

Imperfect geolocation. A potential source of inaccuracy is IP geolocation, as assigning prefixes to a geographic area is challenging and the commercial providers who sell such information use proprietary methods. However, since we have limited our analysis to the country-level, this source of inaccuracy is unlikely to impact CTI, as geolocation databases are typically reliable at that granularity [20, 38, 57]. Further, while determining the location of prefixes originated by large transit providers with a global presence is problematic because of its dynamic nature and wide geographic spread, most networks are much smaller and will have limited geographic presence beyond its primary country of operation [76] (where most or all of its addresses will be located).

We use a country's geolocated IP(v4) addresses as a proxy for the nation's traffic, as this is a limited resource that is necessary to connect any device to the Internet. IP addresses are often used as a proxy for traffic, *e.g.,* in [66], and previous work has found strong correlations between number of IP addresses observed in BGP and traffic volume for ASes that provide either access or transit service [50]. An AS that serves a larger number of IP addresses would consequently have more capabilities for traffic observation.

Finally, we note that although our model has so far only been applied to IPv4 addresses—a reasonable scope given that IPv6 deployment is far from wide in many developing regions, including Africa [14, 49]—the code libraries and software tools we have used are compatible with IPv6, enabling future research in this area.

Inferring Primarily-Transit Countries. Any active campaign launched using publicly available infrastructure will be limited in its effectiveness to reveal peering links by the location of vantage points (VPs) from which the traceroutes are launched. Our campaign is no exception: our VPs are located in a small subset of the world's ASes, and primarily in Europe and North America. However, we argue that our measurements form a sufficient basis to infer that, in the countries we have identified, foreign peering is rare, since: *(i)* we discussed our findings with operators in approx. 10% of these countries, all of whom have confirmed that their nation relies primarily on transit providers to receive traffic from other countries since foreign peering there is rare to nonexistent; *(ii)* while our measurements are launched primarily from the U.S. and Europe, these regions do serve as important content sources and transit hubs (incl. for intracontinental traffic) for countries in Latin America, the Caribbean and Africa [19, 34, 36, 37, 44], where most of the nations we have identified are located.

8 RELATED WORK

Several previous studies have focused on country-level routing, both for the identification of topological bottleneckes [45, 63] and to evaluate the impact of specific countries' ASes on routes towards other countries [43]. All of these studies have used delegation data to map an entire AS to a country; these inferences are prone to inaccuracies when compared with more accurate and granular data such as IP-level geolocation, as important transit ASes may span multiple or many countries, or operate in a country different from their registration. Most recently, Leyba *et al.* [45] addressed the identification of topological bottlenecks, a framework that would also help in quantifying exposure to observation (as CTI aims to address), but with some methodological differences, including: they identify transnational links towards each country using delegation records, and they define bottleneck ASes as those serving the most paths (rather than IP addresses).

Previous work focused on the topologies of specific countries (Germany [72] and China [75]) and relied on country-specific methods and data sets that do not generalize to automatic inference of AS influence in any given country. Fanou *et al.* [32] studied the interdomain connectivity of intracontinental paths in Africa, using a large traceroute campaign (rather than BGP paths).

CAIDA's AS Rank [51] is another topological metric developed to characterize the customer footprint of an AS on the global routing system, *i.e.,* the set of an autonomous system's customer ASes, along with customer ASes of those ASes, recursively. It does not try to capture the capabilities for observation of a transit AS for traffic flowing towards a country; we developed the CTI metric to try to do so.

Our country-level transit influence metric is perhaps most similar to Hegemony [35]. Both metrics aim to identify the transit ASes that are most prevalent on paths towards origin ASes, weighted by the IP address space they serve. Hegemony can be applied either to the global AS-level graph, or to a "Local graph: ... made only from AS paths with the same origin AS" [35]. The latter application is closest to CTI, as this analysis is limited to paths reaching a single origin AS; indeed, we use some of Hegemony local's filtering techniques in our analysis (Sec. 3.2.2). The applicability of (local-graph) Hegemony to the problem of revealing which transit ASes have observation capabilities over traffic flowing towards a specific country—the issue addressed by CTI—is limited, as Hegemony is a metric of centrality of transit ASes on a specific origin AS (not a country). We present a detailed comparision of CTI and (a country-level version of) Hegemony in the countries we study in App. B.

9 CONCLUSIONS AND FUTURE WORK

In this work we tackled the issue of quantifying the exposure of a country's traffic to observation or tampering by specific ASes. The Country-Level Transit Influence (CTI) metric we developed aims to overcome several challenges with making such inferences using BGP data. We apply this metric in a set of countries where transit-provider customer relationships are still an important inbound modality for international traffic, *i.e.,* where international peering is uncommon, a group of nations we identified using analysis of peering databases, our own large-scale measurement campaign, and validation with operators. We applied CTI in these 75 countries to identify the most influential ASes and quantify their international inbound route diversity, and found that the median country has 34% of IP addr. served by a single transit AS.

In the future, we would like to develop measurement and analysis techniques that can be applied to study the exposure of countries that are not primarily served by transit providers, but rather by a dense mesh of bilateral and multilateral peering agreements. We also plan on studying the role of organizations composed of multiple ASes in a country's transit ecosystem (an expansion of our study of state-owned ASes into private companies).

REFERENCES

[1] [n. d.]. AS Rank : About. https://asrank.caida.org/about. ([n. d.]). (Accessed in May 2021).

[2] [n. d.]. Betweenness Centrality - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/betweenness-centrality. ([n. d.]). (Accessed in May 2021).

[3] [n. d.]. Herfindahl-Hirschman Index. https://www.justice.gov/atr/herfindahl-hirschman-index. ([n. d.]). (Accessed in May 2021).

[4] [n. d.]. Internet users - The World Factbook. https://www.cia.gov/the-world-factbook/field/internet-users/country-comparison. ([n. d.]). (Accessed in May 2021).

[5] 2017. Democracy Index 2017: Free speech under attack. https://www.eiu.com/public/topical_report.aspx?campaignid=DemocracyIndex2017. (2017).

[6] 2017. Nauru. Measuring the Information Society Report Volume 2. ICT Country Profiles. https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Nauru.pdf. (2017).

[7] 2019. CAIDA AS-Relationships. http://data.caida.org/datasets/as-relationships/. (2019).

[8] 2019. RIPE Routing Information Service (RIS). https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris. (2019).

[9] 2019. RIR Delegation Files. https://ftp.ripe.net/pub/stats/ripencc/. (2019).

[10] 2019. RouteViews. http://www.routeviews.org/routeviews/. (2019).

[11] 2020. CAIDA's AS-Rank. http://as-rank.caida.org/. (2020).

[12] 2020. Internet Health Report API. Hegemony. https://ihr.iijlab.net/ihr/api/hegemony/. (2020).

[13] 2020. Netacuity. http://info.digitalelement.com/. (2020).

[14] EC Agbaraji, FK Opara, and MI Aririguzo. 2012. IPv6 Deployment Status, the Situation in Africa and Way Out. *International Journal of Advances in Engineering & Technology* 2, 1 (2012), 315.

[15] Aditya Akella, Srinivasan Seshan, and Anees Shaikh. 2003. An empirical evaluation of wide-area internet bottlenecks. In *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement.* 101–114.

[16] Anonymous Authors. 2021. Identifying ASes of State-Owned Internet Operators. In *(Concurrent submission) ACM SIGCOMM Conference on Internet Measurement (IMC '21).* ACM, New York, NY, USA, 13.

[17] Biblioteca del Congreso Nacional de Chile. 2020. FIJA PROCEDIMIENTO Y PLAZO PARA ESTABLECER Y ACEPTAR CONEXIONES ENTRE ISP. leychile.cl/Navegar?idNorma=146170. (2020).

[18] Zachary S. Bischof, Romain Fontugne, and Fabián E. Bustamante. 2018. Untangling the World-Wide Mesh of Undersea Cables. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks (HotNets '18).* Association for Computing Machinery, New York, NY, USA, 78–84. https://doi.org/10.1145/3286062.3286074

[19] Zachary S. Bischof, John P. Rula, and Fabián E. Bustamante. 2015. In and Out of Cuba: Characterizing Cuba's Connectivity. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15).* Association for Computing Machinery, New York, NY, USA, 487–493. https://doi.org/10.1145/2815675.2815718

[20] Bradley Huffaker and Marina Fomenkov and kc claffy. 2011. Geocompare: a comparison of public and commercial geolocation databases - Technical Report. Cooperative Association for Internet Data Analysis (CAIDA), May 2011. (2011).

[21] Xue Cai, M. Rey, CA xuecai, J. Heidemann, CA johnh, and Walter Willinger Niksun. 2013. A Holistic Framework for Bridging Physical Threats to User QoE USC / ISI Technical Report.

[22] CAIDA. 2020. CAIDA Internet eXchange Points (IXPs) Dataset. https://www.caida.org/data/ixps/. (2020).

[23] CAIDA. 2020. Mapping Autonomous Systems to Organizations: CAIDA's Inference Methodology. https://www.caida.org/research/topology/as2org/. (2020).

[24] CAIDA. 2020. Routeviews Prefix-to-AS mappings (pfx2as) for IPv4 and IPv6. http://data.caida.org/datasets/routing/routeviews-prefix2as/. (2020).

[25] CAIDA - PeeringDB. 2020. CAIDA's PeeringDB dumps. http://data.caida.org/datasets/peeringdb/. (2020).

[26] Esteban Carisimo, Julián M Del Fiore, Diego Dujovne, Cristel Pelsser, and J Ignacio Alvarez-Hamelin. 2020. A first look at the Latin American IXPs. *ACM SIGCOMM Computer Communication Review* 50, 1 (2020), 18–24.

[27] Edmond W. W. Chan, Xiapu Luo, Waiting W. T. Fok, Weichao Li, and Rocky K. C. Chang. 2011. Non-cooperative Diagnosis of Submarine Cable Faults. In *Passive and Active Measurement.*

[28] Coral Sea Cable System. 2020. Coral Sea Cable System. https://www.coralseacablesystem.com.au/about/. (2020).

[29] Dainotti, A. and Benson, K. and King, A. and claffy, k. and Glatz, E. and Dimitropoulos, X. and Richter, P. and Finamore, A. and Snoeren, A. 2014. Lost in Space: Improving Inference of IPv4 Address Space Utilization. (Oct 2014).

[30] Amogh Dhamdhere and Constantine Dovrolis. 2008. Ten Years in the Evoultion of the Internet Ecosystem. In *ACM Internet Measurement Conference (IMC).*

[31] Anne Edmundson, Roya Ensafi, Nick Feamster, and Jennifer Rexford. 2018. Nation-State Hegemony in Internet Routing. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18).* Association for Computing Machinery, New York, NY, USA, Article 17, 11 pages. https://doi.org/10.1145/3209811.3211887

[32] Roderick Fanou, Pierre Francois, and Emile Aben. 2015. On The Diversity of Interdomain Routing in Africa. In *PAM.*

[33] R. Fanou, B. Huffaker, R. Mok, and k. claffy. 2020. Unintended consequences: Effects of submarine cable deployment on Internet routing. In *Passive and Active Measurement Conference (PAM).*

[34] R. Fanou, F. Valera, P. Francois, and A. Dhamdhere. 2017. Reshaping the African Internet: From scattered islands to a connected continent. *Computer Communications* 113 (2017), 25 – 42. https://doi.org/10.1016/j.comcom.2017.09.006

[35] Fontugne, Romain and Shah, Anant and Aben, Emile. 2018. The (thin) Bridges of AS Connectivity: Measuring Dependency using AS Hegemony. In *Passive and Active Measurement Conference (PAM).*

[36] Hernan Galperin. 2013. Connectivity in Latin America and the Caribbean: The role of internet exchange points. (2013).

[37] Gustavo Garcia. 2018. Why Miami is Latin America's Center of Interconnection - Interconnections - The Equinix Blog. https://blog.equinix.com/blog/2018/05/01/why-miami-is-latin-americas-center-of-interconnection/. (May 2018). (Accessed on 02/03/2021).

[38] Gharaibeh, Manaf and Shah, Anant and Huffaker, Bradley and Zhang, Han and Ensafi, Roya and Papadopoulos, Christos. 2017. A look at router geolocation in public and commercial databases. In *ACM Internet Measurement Conference (IMC).*

[39] Christopher Giles. [n. d.]. Myanmar coup: How the military disrupted the internet - BBC News. https://www.bbc.com/news/world-asia-55889565. ([n. d.]). (Accessed on 02/05/2021).

[40] Arpit Gupta, Matt Calder, Nick Feamster, Marshini Chetty, Enrico Calandro, and Ethan Katz-Bassett. 2014. Peering at the Internet's Frontier: A First Look at ISP Interconnectivity in Africa. In *Passive and Active Measurement*, Michalis Faloutsos and Aleksandar Kuzmanovic (Eds.). Springer International Publishing, Cham, 204–213.

[41] Infrapedia. 2020. Infrapedia. https://www.infrapedia.com/app. (2020).

[42] International Cable Protection Committee. 2020. Shark Bites and Cable Faults. https://www.iscpc.org/documents/?id=1489. (2020).

[43] Josh Karlin, Stephanie Forrest, and Jennifer Rexford. 2009. Nation-State Routing: Censorship, Wiretapping, and BGP. In *CoRR*, 2009. http://arxiv.org/abs/cs/0608082. (2009).

[44] D. Kiedanski and E. Grampín. 2017. Understanding Latin America IPv6 connectivity: A preliminary exploration. In *2017 36th International Conference of the Chilean Computer Science Society (SCCC)*. 1–6. https://doi.org/10.1109/SCCC.2017.8405137

[45] Kirtus G. Leyba, Benjamin Edwards, Cynthia Freeman, Jedidiah R. Crandall, and Stephanie Forrest. 2019. Borders and Gateways: Measuring and Analyzing National AS Chokepoints. In *COMPASS*.

[46] Liquid Telecom. 2020. Network. https://www.liquidtelecom.com/about-us/our_network. (2020).

[47] Jingang Liu, Wei Peng, Yonglei Yang, and Zhijian Huang. 2013. A Delay-Based Analysis of Multiple Bottleneck Links of End-to-End Paths in the Internet. In *Proceedings of 2013 Chinese Intelligent Automation Conference*, Zengqi Sun and Zhidong Deng (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 93–103.

[48] Shucheng Liu, Zachary S. Bischof, Ishaan Madan, Peter K. Chan, and Fabián E. Bustamante. 2020. Out of Sight, Not Out of Mind: A User-View on the Criticality of the Submarine Cable Network. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. Association for Computing Machinery, New York, NY, USA, 194–200. https://doi.org/10.1145/3419394.3423633

[49] Ioana Livadariu, Ahmed Elmokashfi, and Amogh Dhamdhere. 2018. Measuring IPv6 Adoption in Africa. In *e-Infrastructure and e-Services for Developing Countries*, Victor Odumuyiwa, Ojo Adegboyega, and Charles Uwadia (Eds.). Springer International Publishing, Cham, 345–351.

[50] A. Lodhi, N. Larson, A. Dhamdhere, C. Dovrolis, and k. claffy. 2014. Using PeeringDB to Understand the Peering Ecosystem. In *ACM SIGCOMM Computer Communication Review (CCR)*.

[51] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and Kc Claffy. 2013. AS relationships, customer cones, and validation. In *ACM Internet Measurement Conference (IMC)*.

[52] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, J. Smith, and k. claffy. 2018. Pushing the Boundaries with bdrmapIT: Mapping Router Ownership at Internet Scale. In *Internet Measurement Conference (IMC)*. 56–69.

[53] Alan Mauldin. 2019. Is Your Planned Submarine Cable Doomed? https://blog.telegeography.com/is-your-planned-submarine-cable-doomed. (2019). (Accessed on 02/04/2021).

[54] Babacar Mbaye, Assane Gueye, Desire Banse, and Alassane Diop. 2019. Africa's Online Access: What Data Is Getting Accessed and Where It Is Hosted?. In *Innovations and Interdisciplinary Solutions for Underserved Areas*, Ghada Bassioni, Cheikh M.F. Kebe, Assane Gueye, and Ababacar Ndiaye (Eds.). Springer International Publishing, Cham, 50–61.

[55] Ministry of Posts and Telecommunications. 2010. INTERNATIONAL LONG DISTANCE TELECOMMUNICATIONS SERVICES (ILDTS) POLICY, 2010. http://www.btrc.gov.bd/sites/default/files/ildts_policy_2010_english_0.pdf. (2010).

[56] NAP Chile. 2020. Acerca de NAP. http://www.nap.cl/f_acerca.html. (2020).

[57] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP Geolocation Databases: Unreliable? *SIGCOMM Comput. Commun. Rev.* 41, 2 (April 2011), 53–56. https://doi.org/10.1145/1971162.1971171

[58] Philipp Richter, Florian Wohlfart, Narseo Vallina-Rodriguez, Mark Allman, Randy Bush, Anja Feldmann, Christian Kreibich, Nicholas Weaver, and Vern Paxson. 2016. A Multi-Perspective Analysis of Carrier-Grade NAT Deployment. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. Association for Computing Machinery, New York, NY, USA, 215–229. https://doi.org/10.1145/2987443.2987474

[59] RIPE NCC. 2020. Probes. https://atlas.ripe.net/probes/. (2020).

[60] RIPE NCC. 2020. RIPE Atlas - User-Defined Measurements. https://atlas.ripe.net/docs/udm/. (2020).

[61] RIPE NCC. 2020. RIPE Atlas Probe Archive. https://ftp.ripe.net/ripe/atlas/probes/archive/. (2020).

[62] RIPE NCC. 2020. RIS - RIPE Network Coordination Center. http://www.ris.ripe.net/peerlist/all.shtml. (2020).

[63] Roberts, Hal and Larochelle, David and Faris, Rob and Palfrey, John. 2011. Mapping Local Internet Control. Technical Report, Berkman Center for Internet & Society, Harvard University. (2011).

[64] RouteViews. 2020. Collectors - RouteViews. http://www.routeviews.org/routeviews/index.php/collectors/. (2020).

[65] Anant Shah, Romain Fontugne, and Christos Papadopoulos. 2016. Towards Characterizing International Routing Detours. In *Proceedings of the 12th Asian Internet Engineering Conference (AINTEC '16)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/3012695.3012698

[66] F. Soldo and A. Metwally. 2012. Traffic anomaly detection based on the IP size distribution. In *2012 Proceedings IEEE INFOCOM*. 2005–2013. https://doi.org/10.1109/INFCOM.2012.6195581

[67] Submarine Cable Networks. 2013. MPT, China Unicom Plan International Cable to Boost Internet Connectivity. https://www.submarinenetworks.com/news/mpt-china-unicom-plan-international-cable-to-boost-internet-connectivity. (2013).

[68] Talanei. 2020. ASH Cable buys bandwidth from Tui Samoa. https://www.talanei.com/2018/05/10/ash-cable-buys-bandwidth-from-tui-samoa/. (2020).

[69] TeleGeography. 2020. Submarine Cable Map. https://www.submarinecablemap.com/. (2020).

[70] TM. 2016. TM AND HURRICANE ELECTRIC ESTABLISH STRATEGIC PARTNERSHIP FOR HIGH-SPEED IP BROADBAND SERVICES. https://www.tm.com.my/AboutTM/NewsRelease/Pages/TM-AND-HURRICANE-ELECTRIC-ESTABLISH-STRATEGIC-PARTNERSHIP-FOR-HIGH-SPEED-IP-BROADBAND-SERVICES.aspx. (2016).

[71] TOT Public Company Limited. 2020. TOT: INTERNATIONAL SUBMARINE CABLE. https://www.boi.go.th/upload/content/tot_5d254fe992f21.pdf. (2020).

[72] Wahlisch, Matthias and Schmidt, Thomas and de Brun, Markus and Haberlen, Thomas. 2012. Exposing a Nation-Centric View on the German Internet - A Change in Perspective on AS-level. In *Passive and Active Measurement Conference (PAM)*.

[73] Wired. 2018. What Would Really Happen If Russia Attacked Undersea Internet Cables. https://www.wired.com/story/russia-undersea-internet-cables/. (2018).

[74] A. Zeitoun, Chen-Nee Chuah, S. Bhattacharyya, and C. Diot. 2004. An AS-level study of Internet path delay characteristics. In *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, Vol. 3. 1480–1484 Vol.3. https://doi.org/10.1109/GLOCOM.2004.1378228

[75] Shi Zhou, Guoqiang Zhang, and Guoqing Zhang. 2007. Chinese Internet AS-Level Topology. *IET Communications* 2, 1 (April 2007).

[76] Ran Zhuo, Bradley Huffaker, kc claffy, and Shane Greenstein. 2021. The Impact of the General Data Protection Regulation on Internet Interconnection. *Telecommunications Policy* 45 (2) (2021).

A VALIDATION

In this section we describe our discussion of CTI findings with operators, as well as an analysis of CTI's temporal stability. We also summarize our discussions with: *(i)* operators regarding our inferences of transit-dominant countries, *(ii)* ASes with prefixes geolocated to these countries. *Our findings are largely consistent with each operator's view of the transit ecosystem of the countries discussed with them*: we did not find any false positives in our identification of primarily-transit countries, and the per-country rate of true positives—in terms of influential transit ASes confirmed by the operators in 6 countries—was 83%, on average[12].

A.1 Operator Validation

We discussed our findings with employees or contractors of two types of organizations: commercial network operators and non-profits who conduct networking research (universities, registrars, and non-commercial network operators). Additionally, we describe the results of our discussions following a mass email request to ASes with prefixes geolocated in countries in our study. Discussions with all but one of these organizations are anonymized following their requests. These discussions took place in the spring of 2020, unless otherwise specified.

A.1.1 Commercial Network Operators. We emailed a former and 8 current employees at 9 companies operating transit and/or access networks primarily in Africa and Latin America. An operator confirmed that they operate a large transit network in two specific countries in Sub-Saharan Africa. They also confirmed that one of those countries is primarily-transit when asked directly about that specific country. Two operators in Africa (one of which is Liquid Telecom[13], the sole non-anonymized conversation we report) broadly confirmed being a transit provider for traffic flowing towards the countries we indicated, but refrained from confirming that the countries as a whole are primarily-transit. One former and one current operator in a single Sub-Saharan African country confirmed that its inbound traffic flows primarily through transit links. We sent these operators the set of top 7 ASes[14] by CTI in that country. An operator responded with four ASes that they state are the only direct upstreams of a large access network; these are ranked 1, 2, 5 and 6 by CTI in that country. The second operator confirmed the top 5 of the 7 we sent; these are ranked 1, 3, 4, 5 and 6 by CTI in that country. One operator in LACNIC confirmed—in Oct. 2020—that their country is primarily-transit. They also confirmed 12 of the 15 top ASes identified by CTI in that nation as influential operators. Two additional operators never responded, while a third one declined our request.

A.1.2 Networking researchers at non-profits. We contacted 16 researchers in 10 countries in Africa and Latin America. In 5 of those countries (four in Africa, one in Latin America), 6 researchers confirmed the countries are primarily-transit; of these, one (from Latin America) declined to comment on the list of top ASes we sent them, while the discussions in that regard with the other five are described in the next paragraphs. Two other researchers declined to comment altogether, and 8 did not respond. Two researchers in two different Sub-Saharan African countries confirmed 8 of the 10 ASes in the top 10 we sent[15]. A researcher in a country in Sub-Saharan Africa confirmed 5 of the top 7 ASes by CTI[16]. Two researchers in a single country in North Africa responded to our set of top ASes by CTI. The first researcher was able to confirm that 7 of the 10 ASes are transit providers of access ASes operating in this country. The second researcher's response in this regard was to dispute an AS in the top 10 by CTI, and suggesting that we investigate the transit providers of the country's access networks[17]; the disputed AS is an inferred transit provider of one of those access networks, a relationship which was directly confirmed by the first researcher.

A.2 ASes with Prefixes Geolocated

We sent a mass email request to the WHOIS abuse address registered by ASes that had prefixes geolocated in 10 countries[18] (with IRB approval): BO, CO, VE, CM, BD, GT, CL, HN, SV and ZW. These were selected as a mix of large and

[12] The per-country true positive rates are, sorted increasingly: 66%, 70%, 80%, 90%, 90%, 100%.

[13] Zimbabwe, Zambia, Lesotho, Somalia, D.R. Congo.

[14] Unfortunately due to the timing of the validation process, we sent a set of ASes—we did not include actual CTI values in the message, just the set of top ASes—to these operators that was produced before updating our CTI methodology to its current form; 6 of 7 ASes are present in both our final CTI top 7 and in the outdated list we sent them.

[15] See previous footnote; in this case 9/10 ASes remained constant across both sets for both countries, including the 8 the researchers confirmed. The operators also confirmed an additional AS from the outdated set in each country, which are ranked 12 and 13, respectively, in the final CTI tally.

[16] See previous footnotes. In this case 5/7 ASes are constant across both sets; the operator confirmed two additional ASes from the outdated set, which are ranked 9th and 10th by CTI in the final tally.

[17] Which they listed in their response; we found that these networks originate 99.7% of addresses in that country, an anecdotal but encouraging sign regarding the correctness of our assessment of which ASes originate addresses in this particular country.

[18] We only contacted ASes who had at least 1% of their addresses in the country, and since this survey took place in 2021, we use the addresses geolocated in Jan. of that year.

small (by #ASes) countries where English or Spanish are among the primary languages. We received 111 responses in 9 of these countries (all but ZW). Of these, 107 confirmed they operate primarily in the country that we geolocated their prefixes to[19]. Additionally, 108/111 were willing to discuss which type of business relationship dominated their inbound international traffic. Of these, 83 stated that transit relationships are the primary modality.

B COUNTRY-LEVEL HEGEMONY

In this appendix, we build a country-level alternative metric based on Hegemony [35] and compare CTI to it. The reason for the comparison is to determine if CTI is too aggressive in its filters, discarding too much input data. For that purpose, we build a benchmark using local hegemony, a metric of centrality of any AS (including both transit providers and peers) on paths towards a single origin. Hegemony consists mostly of a single filter on input BGP data, making it an appropriate benchmark. This benchmark was not trivial to build, as hegemony local produces a bilateral metric of influence between a transit AS and an origin AS on the global topology. We find that both metrics tend to agree about the country-level influence of marginal ASes and very dominant ASes (these two groups include the vast majority of ASes), but they diverge among providers with high influence that are not dominant. This last group is among those we aim to discover with CTI, on which our heuristics intend to operate. Hence, we find that out heuristics are not overly aggressive nor unduly excluding input data.

No existing metric provides an assessment of a country's exposure to external observation as CTI does. In order to quantify the impact of our heuristics on the resulting CTI, we built an alternative country-level metric based on AS Hegemony [35] as applied to a local (*i.e.,* all paths originated by the same AS) graph. This analysis gives us an indication of how our heuristics and filters compare to those of a conceptually equivalent alternative.

While Hegemony is concerned with extracting the most accurate estimate of centrality on an existing graph, and not with estimating country-level inbound route diversity as CTI, it is possible to build a metric that serves a similar purpose as CTI, which we call *country-level hegemony* (*CLH*) as follows:

$$CLH(AS_t, C) \in [0,1] = \sum_{AS_o \in (C)} H(AS_t, AS_o) \cdot \frac{a^*(AS_o, C)}{A(C)}$$

where $H(AS_t, AS_o)$ is the hegemony score of AS_t on AS_o during the same period[20] in March 2020 when we applied CTI, (all the other terms have been previously introduced in Eq. 5.3).

In other words, CLH is a conceptually equivalent metric to CTI. For each AS-country pair (a transit AS serving a country) in our study of 75 countries, we show both CTI and CLH in Fig. 9. These metrics tend to agree on a score for a given transit AS in a given country: a linear regression has a slope of 0.9988, intersection of 0.002, and R^2 of 0.87. The takeaway is that the heuristics of CTI do not introduce unnecessary noise to our analysis because, on aggregate, a country-level alternative based on Hegemony—which applies a single filter to BGP data in order to estimate AS centrality, and excludes considerably fewer BGP monitor than CTI does—tends to agree with CTI's assessment.

Further, the metrics tend to produce qualitatively similar assessments of each transit AS in each country: they either assess the AS as marginal—both metrics assign it a score lower than 0.1 (6,124 of 6,428 AS-country pairs); or they assess the AS as very dominant—both assign a score greater than 0.6 (6 AS-country pairs).

There is, however, disagreement among the metrics in the middle section of Fig. 9 (remaining 298 AS-country pairs); we study the data points in the area where either metric has a score between 0.1 and 0.6). In particular, we investigate the data points where the two metrics disagree the most in the remainder of this section.

We manually inspected 15 AS-country pairs where CLH produces a much higher score than CTI ($|CLH - CTI| > 0.2$), or viceversa. We stress that the true score for each AS-country pair is unknowable, so what we intend to evaluate is the impact of the individual components of CTI on these data points. To that end, we compute an alternative set of CTI scores, where the indirect transit discount (a coarse heuristic defined in Sec. 3.1.1) is not applied.

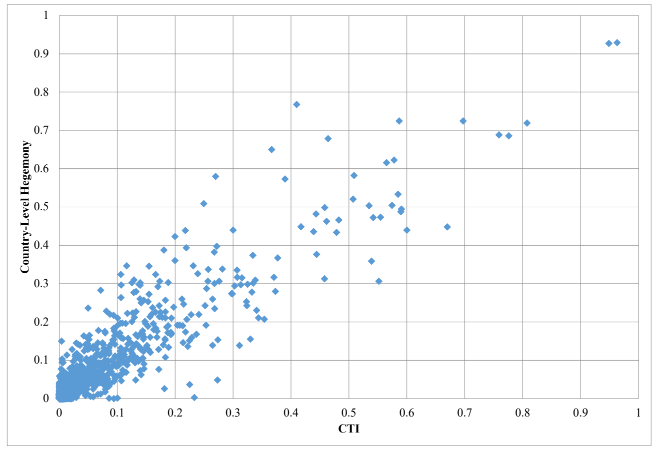


Figure 9: CTI and CLH scores for the 6,428 AS-country pairs in our study.

[19]In 3 cases, they stated that the country was among their primary places of operation, but that they also operated in other countries.

[20]As Hegemony is published in 15-min intervals [12], we take the 5-day average score.

For the 11 AS-country pairs where CLH produces a higher score than CTI, the indirect transit filter has a meaningful impact on our estimate of CTI: not applying this filter would have increased the CTI score by a median and average of 93% and 102%, respectively. Indeed, for these 11 AS-country pairs, the indirect transit filter causes 95% (median) and 96% (metrics) of the gap between the metrics. Since our purpose was to produce a conservative estimate of the transit influence of indirect transit providers, we find that this CTI heuristic is working as intended.

All AS-country pairs where CLH produces a meaningfully lower score than CTI measure influence on island nations: Nauru, Samoa, East Timor, and St. Vincent & Grenadines. The indirect transit filter has a minor impact on CTI for these 4 AS-country pairs: between 0-2.5%. The disagreement between the metrics, then, primarily stems from the core analysis each does over BGP paths. One of these AS-country pairs involve C&W (a previously introduced influential AS in the Caribbean, Sec. 4.1), an operator which owns or co-owns submarine cables landing in St. Vincent & Grenadines, which suggests that the operator is likely influential on these island, potentially justifying a high CTI score. One other AS-country pair relates to a small island in the South Pacific, Nauru, reportedly relying primarily on satellite connections for international connectivity [6]. Estimating inbound route diversity in this nation may be particularly challenging for any metric, but they are nonetheless likely exposed in their external connectivity (and therefore their inclusion in our study is justified given our goal of identifying exposed nations).

C BGP MONITOR LOCATION

We begin with the 685 monitors in RIPE and RouteViews. We discard (91) monitors aggregated at multi-hop collectors and monitors that are not full-feed, so we are left with 350 monitors in 209 ASes. We determine the location of each full-feed BGP monitor as follows. First, we find the locations of RouteViews and RIPE RIS BGP collectors. We build a first set of locations by finding RIPE Atlas probes co-located at Internet Exchange Points (IXPs), by searching the list of peers for the IXP name, and assign that probe to the country where the (single-location) IXP is present, *e.g.,* BGP RRC01 – LINX / LONAP, London, United Kingdom. We confirm the BGP monitor location by running `ping` measurements from RIPE Atlas probes hosted at the IXP to the BGP monitor's IP address, and conclude that the BGP monitor is in the same city as the IXP if the RTT is lower than 5 ms. For the remaining BGP monitors we look for available RIPE Atlas probes in the ASes that peer with the same BGP collector, and similarly run `ping` measurements towards both the BGP monitor's IP address and a RIPE Atlas probe located in the same city as the one listed for the monitor. We conclude that the BGP monitor and RIPE Atlas probe are in the same city if both sets of RTTs are under 5 ms.

We exclude 118 monitors at this stage because there is no available RIPE Atlas probe hosted at the IXP (in the city where the monitor is listed) nor at any of the other peers of the collector aggregating announcements from the BGP monitor. We discard remote peers from our set, those that have `ping` RTTs higher than 30 ms from the RIPE Atlas probe in the BGP monitor's listed city. For monitors with an RTT between 5–30 ms, we infer them to be at the listed location if we get confirmation using DNS records—*i.e.,* we find a geographical hint such as a three-letter city or airport code, or the full name of the city, using a reverse lookup with the BGP monitor's IP address—or a matching country of the BGP monitor's `peer_asn` record in the RIPE RIS or RouteViews collector list [62, 64]. Our final set M has 214 monitors in 145 ASes and 19 countries. We quantify the aggregate impact of all of our filters, including the exclusion of certain BGP monitors per country, in App. B, and find that their impact is within reason given an alternative metric built using previous research [35].

D CTI DERIVATION

A potential metric needs to evaluate the frequency at which a specific transit provider appears along paths towards an origin (destination) AS_o. Mathematically, for a transit AS_t serving inbound international traffic to any origin AS_o in country C from any third AS_n, AS_t's centrality, ASC, is defined as

$$ASC(AS_t, C) \in [0,1] = \frac{\sum_{AS_o \in C} S(AS_t, AS_o)}{\sum_{AS_n} S(AS_n, AS_o)}, \qquad (2)$$

where $S(AS_t, AS_o)$ is the number of paths towards AS_o where AS_t is present as a transit provider, and $S(AS_n, AS_o)$ is the total number of inbound paths towards AS_o from transit AS AS_n. This definition of ASC is directional (paths going from outside the country towards the origin in the country) and excludes paths where no p2c relationship is present (*i.e.,* peering paths) as we perform our evaluation of transit influence only in countries where visible foreign peering is not prevalent.

There are two major shortcomings of the model expressed in Eq. 2 for our purposes: first, it provides no mechanism to compute AS centrality on a country, *i.e.,* a collection of IP addresses originated by multiple ASes. Second, Eq. 2 treats paths towards any AS_o equally, regardless of how many IP addresses each of them originates. Since we are interested in measuring the exposure to observation of a country's address space (*not* origin ASes), we adapt Eq. 2 to quantify the IP-level centrality (IPC) of a transit AS AS_t along paths towards addresses originated in each nation C as

$$IPC(AS_t, C) \in [0,1] = \sum_{p|\text{onpath}(AS_t,p)} \frac{a(p,C)}{A(C)}, \quad (3)$$

where onpath(AS_t, p) is true if AS_t is present on a BGP path towards a prefix p, $a(p,C)$ is the number of addresses in prefix p geolocated to country C, and $A(C)$ is the total number of IP addresses geolocated to country C. Note that this is a departure from established models of AS centrality, such as Betweenness Centrality [2]. This is because the leaves of the graph are geographically-annotated nodes, or a single origin AS abstraction per country representing the union of each of the origin ASes' addresses present in the country. Further, IP addresses determine the weight of the edges terminating in them, and the core of the graph being ASes who carry traffic towards the edge only. In practice, a single AS may serve as an origin AS and a transit AS in the same country, a case in which our model creates two abstractions for the same AS: one as a transit provider for other origin ASes, and the other as a component of the country's compendia of originated addresses.

In theory, IPC is a measure of the IP-level centrality of an *AS*, with the highest values of IPC for any set of transit ASes serving a country providing an indication of how exposed to observation that nation's inbound routing ecosystem is. A country where many transit ASes have similar values of IPC (for example, 10 ASes with $IPC = 0.1$) will likely be less exposed than another nation where an AS has high IPC (for example, a single AS with $IPC = 1.0$) and all other transit ASes are marginal ($IPC \leq 0.01$). Therefore, comparing the distribution of IPC values across countries gives us an indication of the relative resiliency of a country's inbound routing infrastructure compared to other nations.

In practice, IPC is the core term (but not the only component) of CTI in Eq. 1, the metric we use to compute the influence of any given transit AS_t on a graph composed of inbound paths towards a country C. The remaining terms in that equation are necessary to account for the shortcomings of BGP data, the most important of which we now describe. In order to compute CTI, we rely on the largest compendia of publicly-available BGP routing data, collected by RouteViews [10] and RIPE RIS [8], who aggregate BGP messages from operational routers (BGP monitors) at cooperating ASes. CTI applies several analysis techniques (Sec. 3) to infer IPC from these BGP measurements and address a core technical challenge: BGP data collection is heavily biased towards paths seen from the (small sample of) ASes hosting BGP monitors. As monitors are not distributed uniformly across and between countries, and many countries and most ASes have none, the inferences of transit influence built with these measurements will heavily oversample paths towards ASes hosting BGP monitors.

Because no ground truth of a country's inbound route diversity exists, and backup or less preferred links that are only announced in the presence of routing disturbances are therefore unobservable most of the time, our techniques include coarse heuristics (Sec. 3.1.1) based on a conservative guiding principle: in parts of the graph with the most uncertainty—*i.e.,* of nodes not directly connected to the origin—we artificially discount transit influence estimates. The uncertainty increases with AS-level hops from the origin, as each additional AS on the path may have unobserved alternative transit providers (See Fig. 2). This admittedly conservative approach allows us to compute a lower bound of country-level observation exposure in cases where indirect transit providers appear along paths towards a large fraction of a country's addresses. At the same time, this heuristic does not impact our assessment of route diversity resulting from influential ASes in close topological proximity to the origin (an area of the graph where there is less uncertainty).

E SUBMARINE CABLE ASES

Because submarine cable operators are inherently linked to physical infrastructure, it is possible to construct an alternative view of the findings in Fig. 4 based on the actual submarine cables being operated. If a single cable is linked to an AS with high CTI in multiple countries, an event affecting that cable (which may include weather-related or "anchoring and fishing activities" [42], as well as targeted attacks) may have serious consequences in multiple countries [73]; such cables and the associated ASes are listed in Table 3. Most notably, C&W (formerly Columbus Networks) is among the top providers in 10 countries in Central America and the Caribbean thanks to its ownership of the ECFS and ARCOS-1 cables. Its CTI in those countries ranges from transit dominance in small islands, to a more marginal position in Central America. Tata, Telefonica and Bharti Airtel also have an important transit presence in West Africa, Western South America, and South Asia respectively. Table 3 also highlights the critical nature of the SeaMeWe family of submarine cables, with top ASes by CTI being identified as a co-owner of one of the cables in eight countries. Cables in this family have received attention in previous studies [27].

Table 3: Submarine cables and their top AS operators by CTI. ASes listed match the countries from left.

Sub. Cable	ASes (# of countries, if more than one)	Countries
SeaMeWe-4	9498-BHARTI Airtel (2), 8452-TE	India, Bangladesh and Egypt
EIG	37558-LIT, 9498-BHARTI Airtel, 8452-TE	Libya, India, Egypt
SeaMeWe-5	45489-SL Tel., 9121-Turk Tel., 8452-TE	Sri Lanka, Turkey, Egypt
AAE-1	15412-Reliance, 8452-TE, 8781-Ooredoo, 38040-TOT, 8529-Omantel	Yemen, Egypt, Qatar, Thailand, Oman
EASSy	16637-MTN, 37662-WIOCC	Sudan, Somalia
SeaMeWe-3	6762-TIS, 8452-TE, 45558-MPT, 9121-Turk Tel., 6939-HE[1]	Morocco, Egypt, Myanmar, Turkey and South Korea
ACE	327903-Ministry I&C, 37529-GITGE, 8346-Sonatel, 29544-Mauritanian Tel.	Sierra Leone, Eq. Guinea, Guinea, Mauritania
WACS	6453-TATA, 30844-Liquid Tel., 15964-Camtel	Cape Verde, Congo DRC, Cameroon
ECFS	23520-C&W (5)	S.V.G., S.K.N., St. Lucia, Barbados, Trinidad & Tobago
ARCOS-1	23520-C&W (5)	Honduras, Nicaragua, Venezuela, Belize, Bahamas
Pan-Am	12956-Telefonica (3)	Ecuador, Peru, Bolivia[2]
AMX-1	14754-Telgua, 14080-Telmex Colombia	Guatemala, Colombia

[1] Partnership with Telecom Malaysia [70].
[2] Not included in Fig. 4 as it is landlocked.