CS 8803 Internet Data Science

Alberto Dainotti

1/18/2022

Georgia Tech

Welcome to Day 11

Today's paper

Published at IMC 2018

Presented in classroom by Alberto Dainotti

Thanks to Zachary Weinberg for sharing their slides

How to Catch when Proxies Lie

Verifying the Physical Locations of Network Proxies with Active Geolocation

Zachary Weinberg · Nicolas Christin · Vyas Sekar
Carnegie Mellon University

Shinyoung Cho *SUNY Stonybrook*

Phillipa Gill *UMass-Amherst*

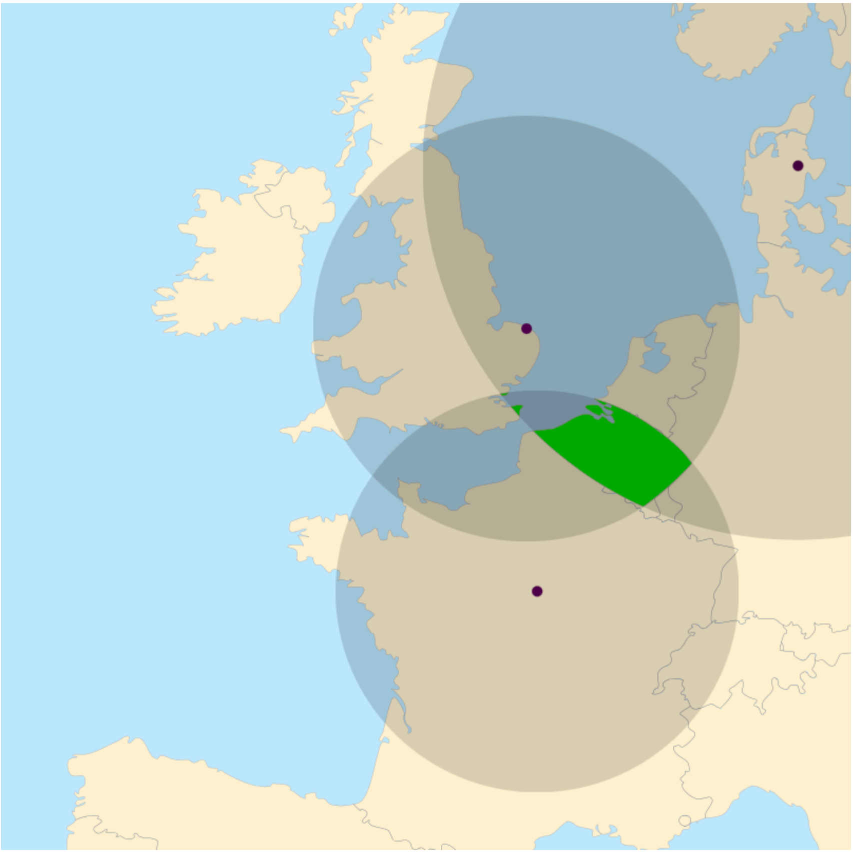
1 / 25

Background: IP Geolocation

- Methods for finding the physical location of Internet hosts
- Passive:
  - Location information from Regional Internet Registries
  - Location encoded in router hostnames
  - Private consultation with individual ISPs
- Active:
  - Measuring RTT between *landmark* hosts and a *target* host
    - If the target is very close to a landmark then we can approximate its location with the landmark's
    - More commonly we use *multilateration…*

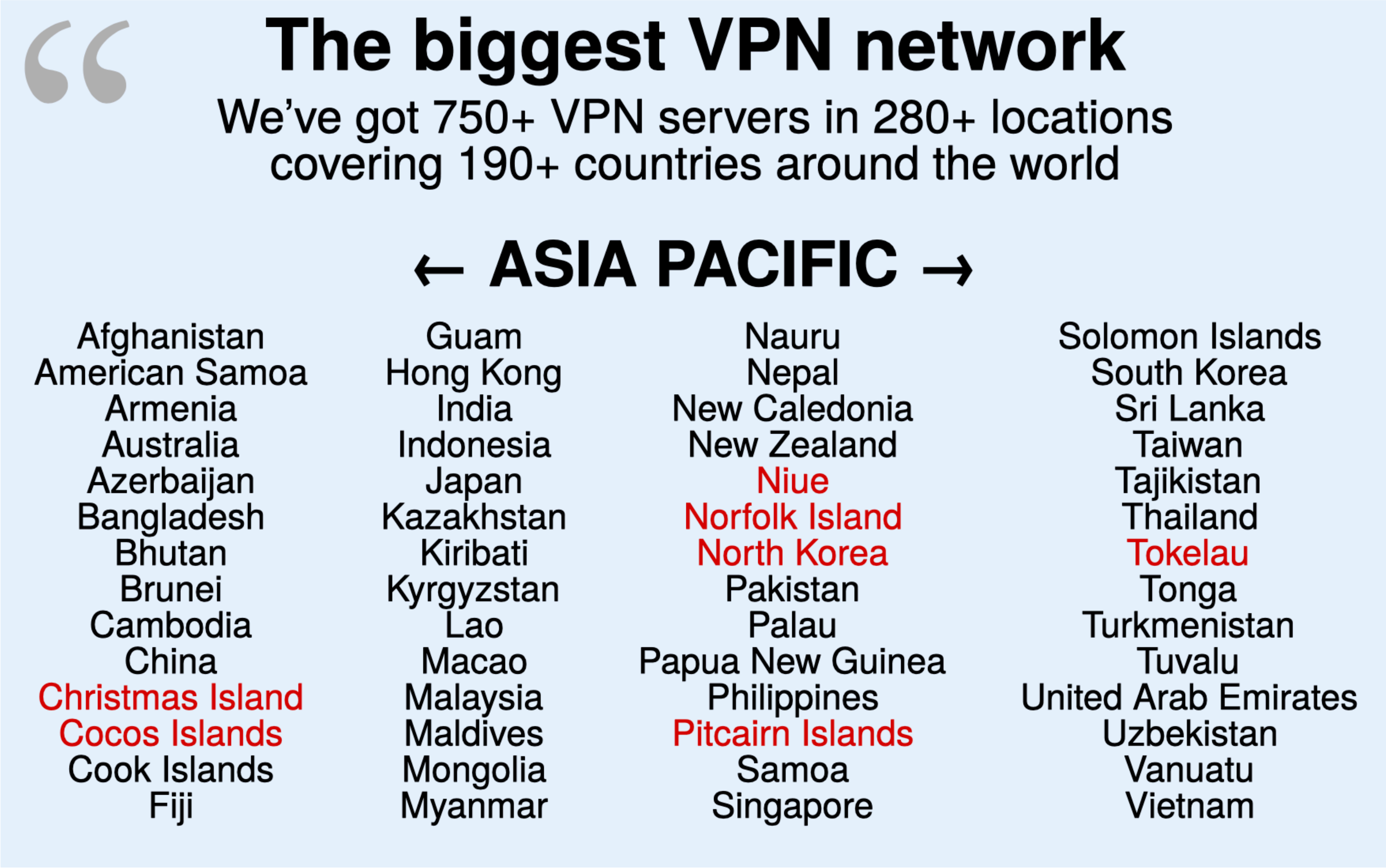
Background: IP Geolocation: *Multilateration*

- Given a *target*, for each *landmark* we estimate the maximum distance that a packet could have traveled (from the landmark to the target—and back) in the time measured
  - Draw disks on a map bounded by these distances
  - The target must be where all the disks intersect
- Same principle as GPS
- Problems with multilateration in IP Geolocation
  - Packets do not travel in straight lines
    - Cables are laid on practical paths
    - Network routes are optimized for bandwidth, not latency and are based on economic relationships between ASes
    - Intermediate routers can add unbounded delays
  - Research on models for delay-distance relationship



Paper summary

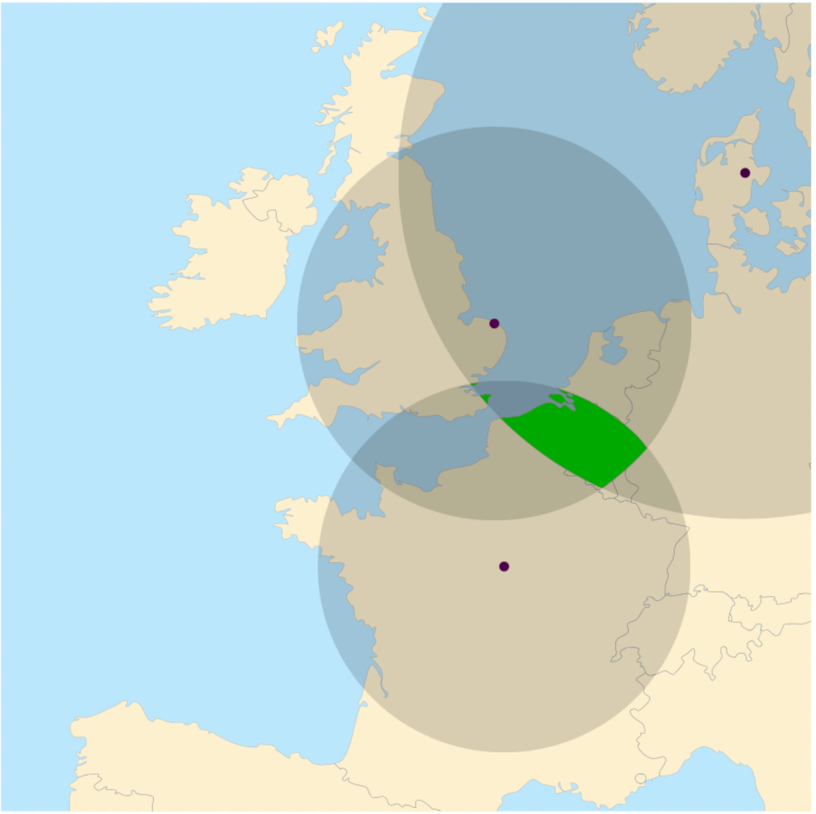
- VPN providers compete by speed, privacy, breadth of ***locations***
- In this paper they apply active geolocation to check advertised locations of VPN servers
    - Can locate a VPN server within 1000km2 radius. Enough to disprove claims
- Adapt + Improve active geolocation
- Finding: at least a third of all the servers they tested are not in the advertised country.

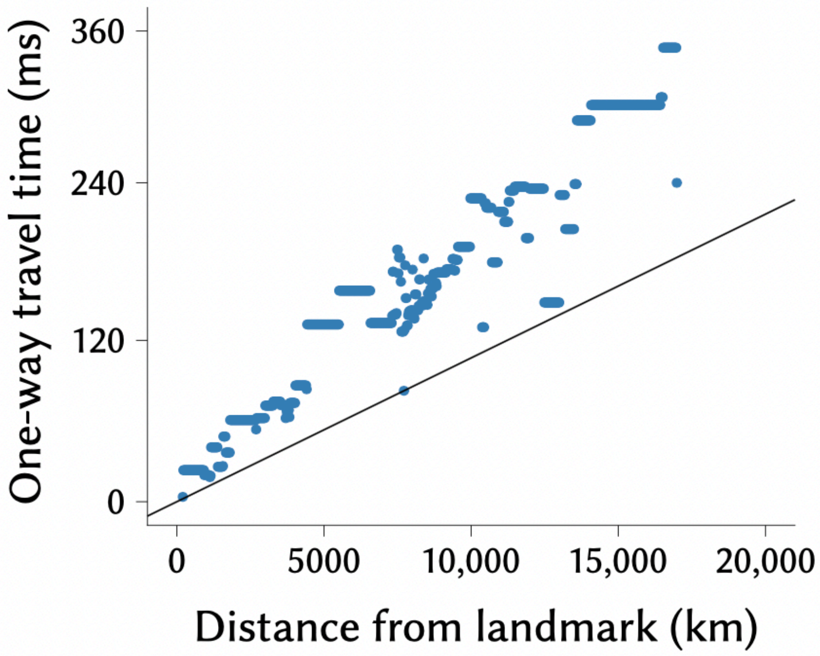


Challenges

- There are different techniques/algorithms. Not clear what's the best
  - Select/evaluate algorithms
  - Not much implementations available
- Need for landmarks
  - Increasing number of landmarks improves accuracy but slows down the measurement process
- Need for validation
  - Crowdsourcing
- Proxies (VPN servers) typically don't respond to probing.
  - We can only send packets *through* the proxies

Active geolocation

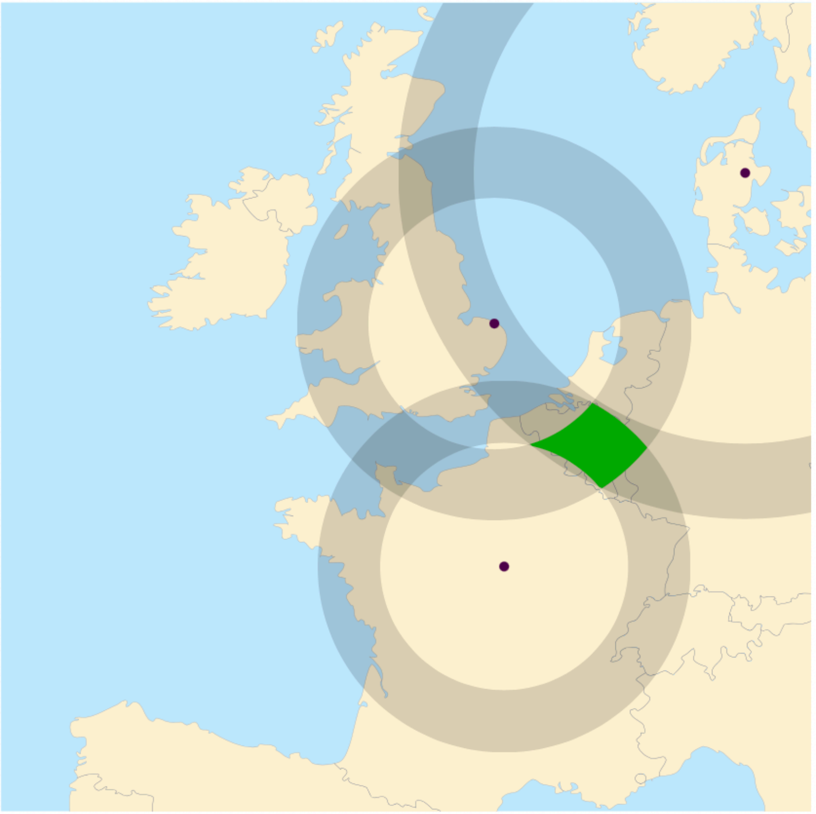


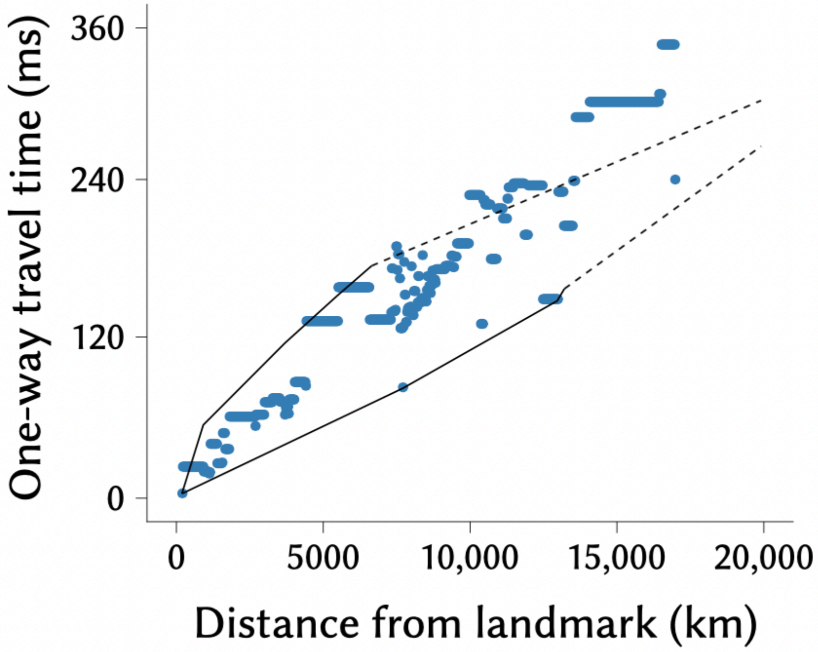


Same principle as GPS, but use packet round-trip time (RTT)

CBG: Linear estimate of maximum packet travel

(Quasi-)Octant

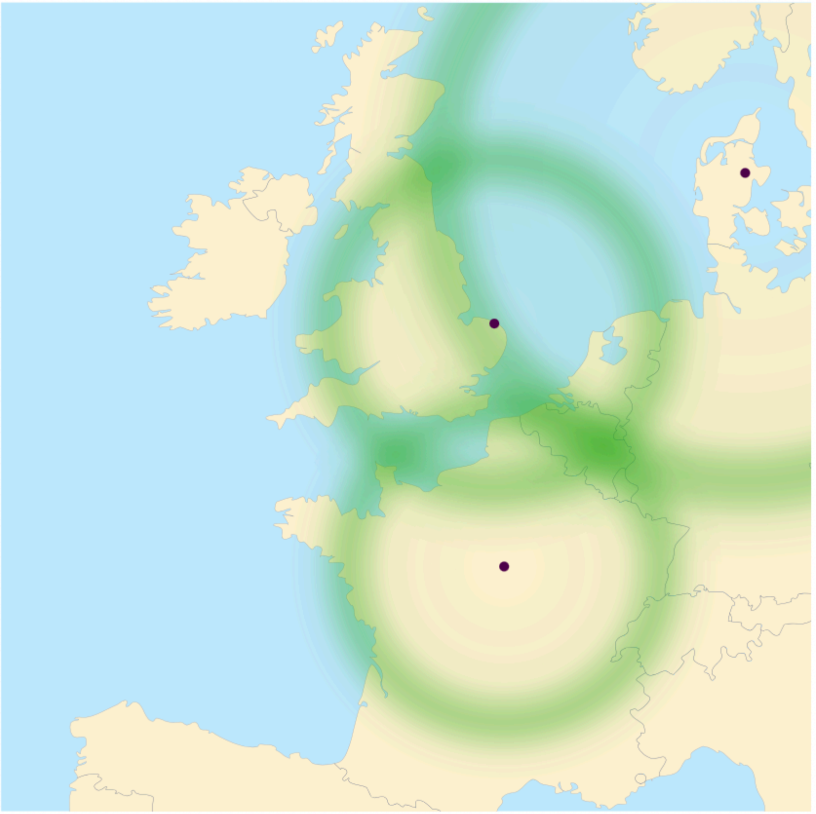




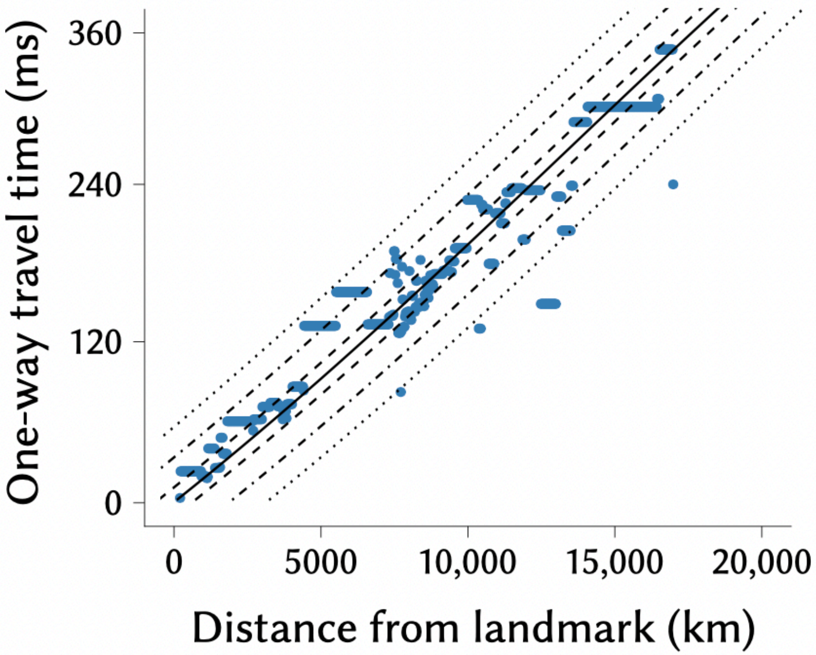
Minimum as well as maximum distance

Piecewise-linear travel time estimate, using convex hull of points

Spotter



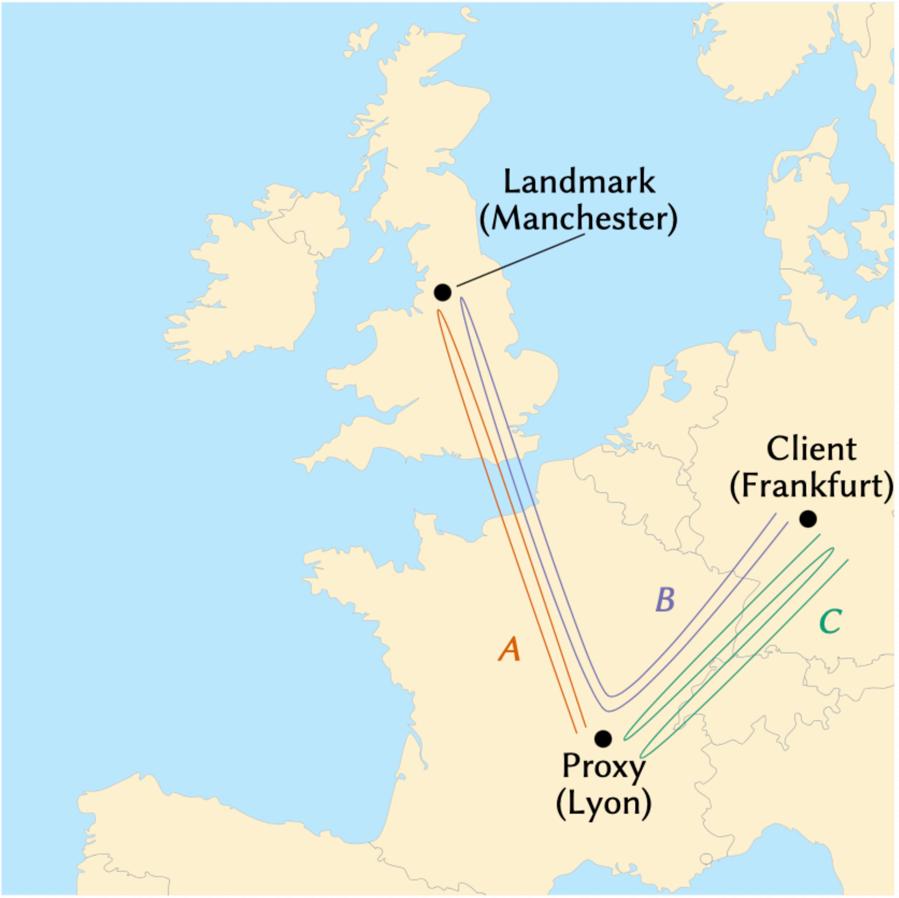
Probabilistic combination
of Gaussian rings



Cubic polynomial
estimates of μ and σ

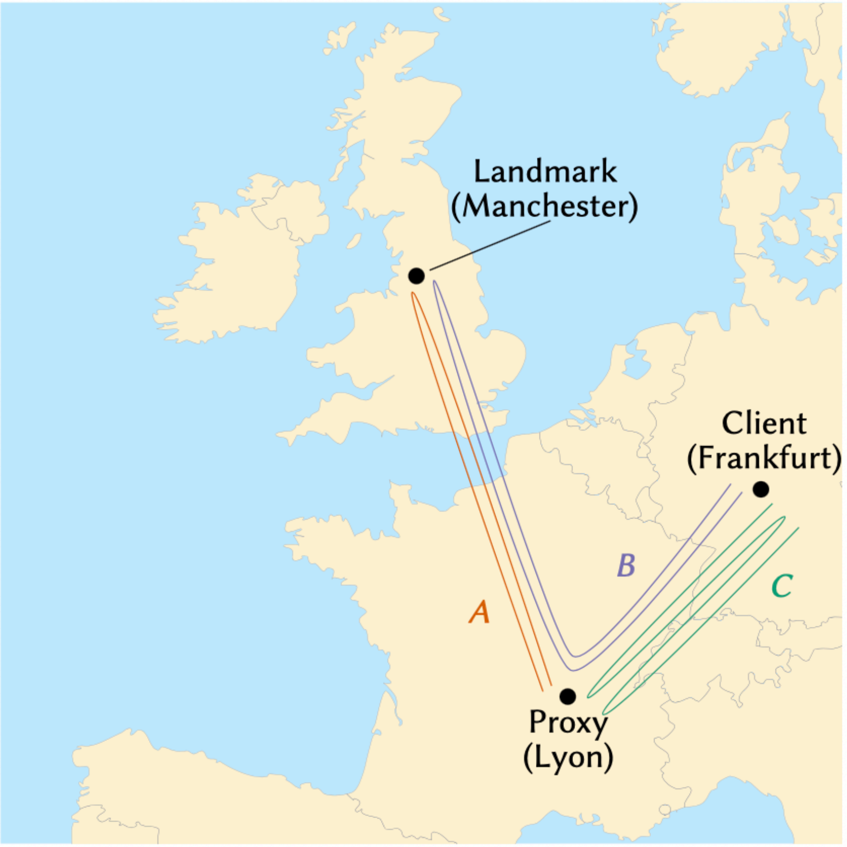
Measurement approach: adaptations for proxies

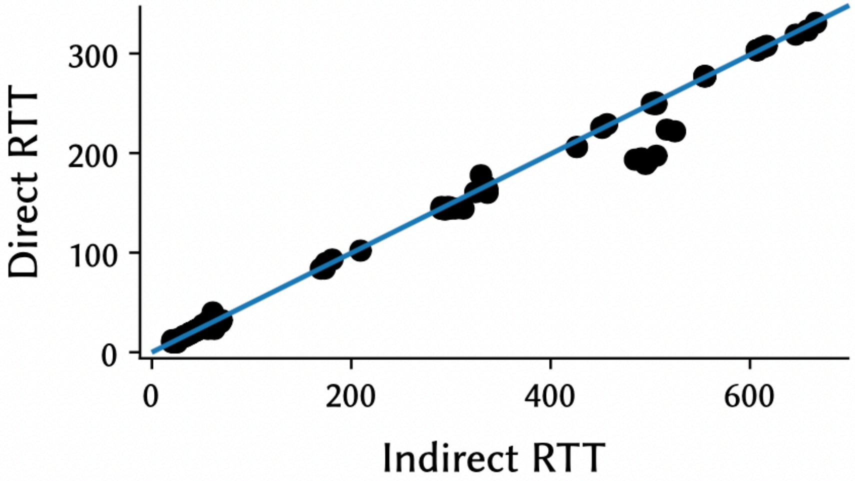
- We cannot ping the proxy
- We need to go through them
- *RTT measured = (<u>RTT from client to proxy</u>) + (RTT from proxy to target)*
  - B = 0.49*C + A
- We have the client ping itself through the VPN



Cannot measure *A*
Can measure *B* and *C*

Measurement through VPN servers





This is based on proxies that can be directly pinged!

Cannot measure A
Can measure B and C

$$A = B - 0.49C$$

Measurement approach

- Implement and test 4 pre-existing techniques
  - *https://github.com/zackw/active-geolocator*
- Test/validate them and add a 5th one that is an improvement

- Landmarks: used RIPE Atlas *anchors* and *probes*
- To speed up the process / reduce the number of landmarks we use a two phase measurement (Sec. 4.1)
  - We first measure RTTs to 3 anchors per continent, and use these measurements to deduce which continent the target is on
  - We then randomly select and measure RTTs to 25 more landmarks on that continent
- …

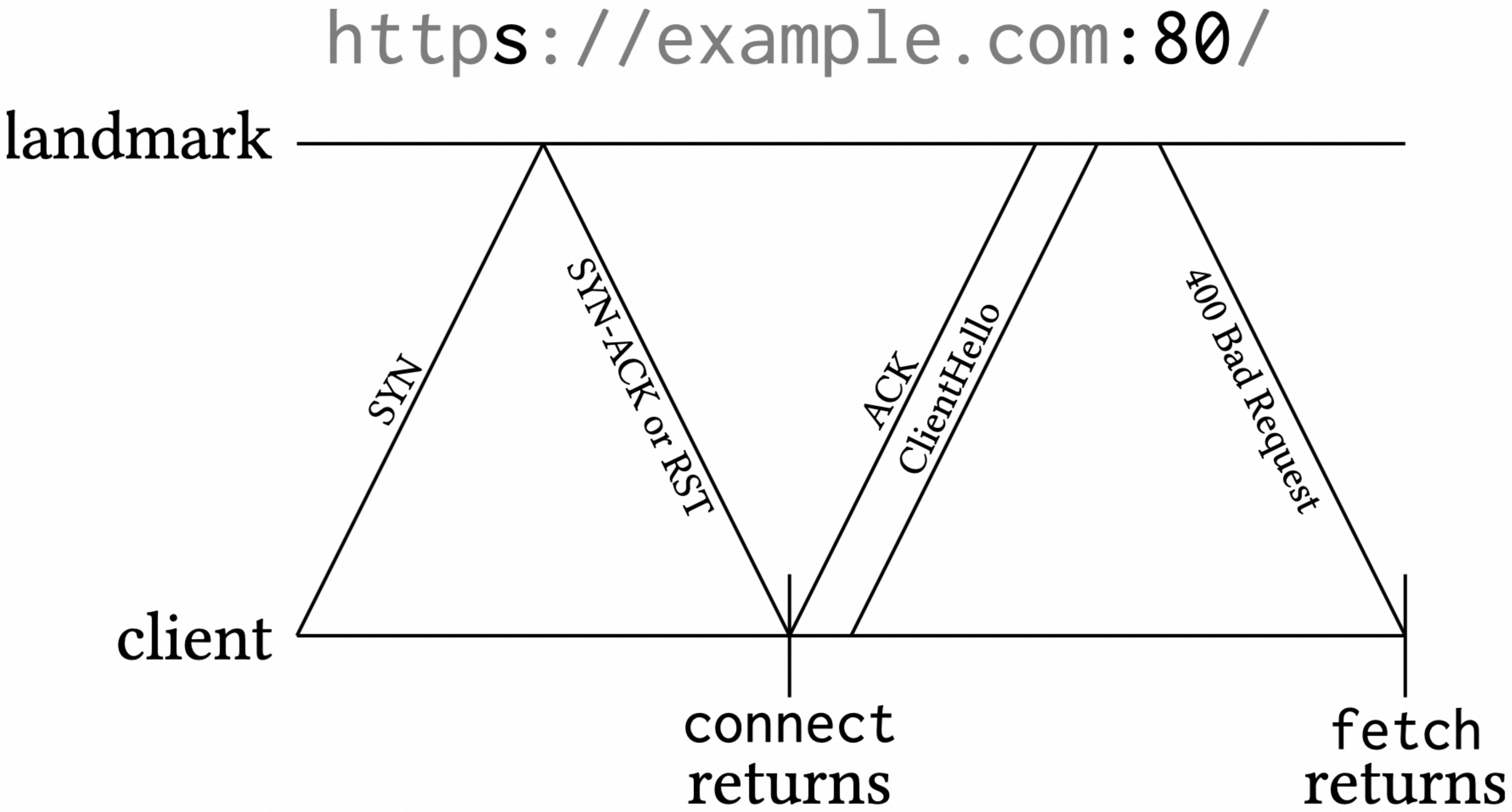
Measurement approach: maintain a server/list/models

- We maintain a server that
  - retrieves the list of anchors and probes from RIPE's database every day
  - selects the probes to be used as landmarks
  - and updates a delay-distance model for each landmark, based on the most recent two weeks of ping measurements available from RIPE's database
- Our measurement tools retrieve the set of landmarks to use for each phase from this server, and report their measurements back to it
-

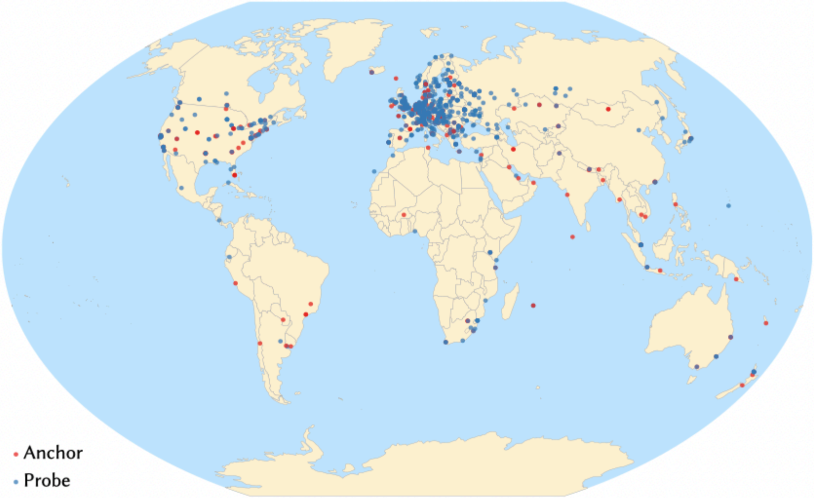
Measurement approach: *connections*

- The only type of network message we can reliably use to measure round-trip time is a TCP connection on a commonly used port, e.g. 80 (HTTP).
- We implemented two measurement tools that use this method to measure round-trip times to each landmark.
  - Command line
    - Attempts a TCP connection from the client to the landmark -> obtains RTT between proxy & landmark
  - Web-based (https://research.owlfolio.org/active-geo/)
    - Website hosting the application
    - Runs on the browser
    - More complicated and more uncertainties
- Why need for also a web-based version?
  - To validate on a global unbiased sample

Measuring RTT with a Web app

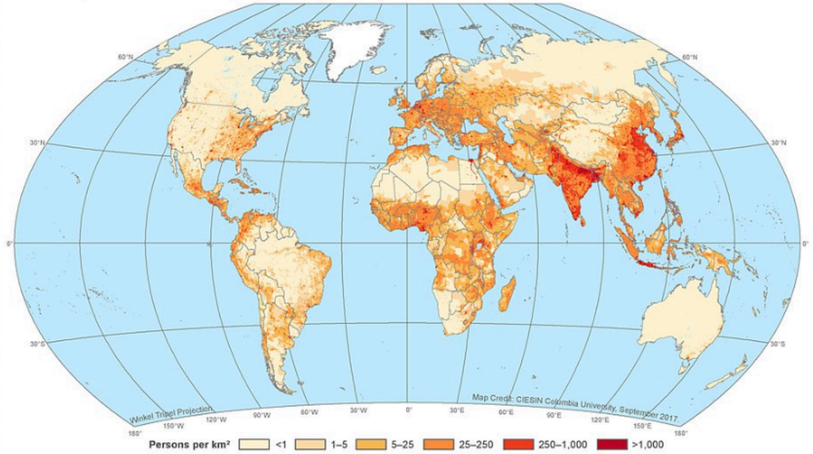


Testing active geolocation around the world



RIPE Atlas anchors
and stable probes

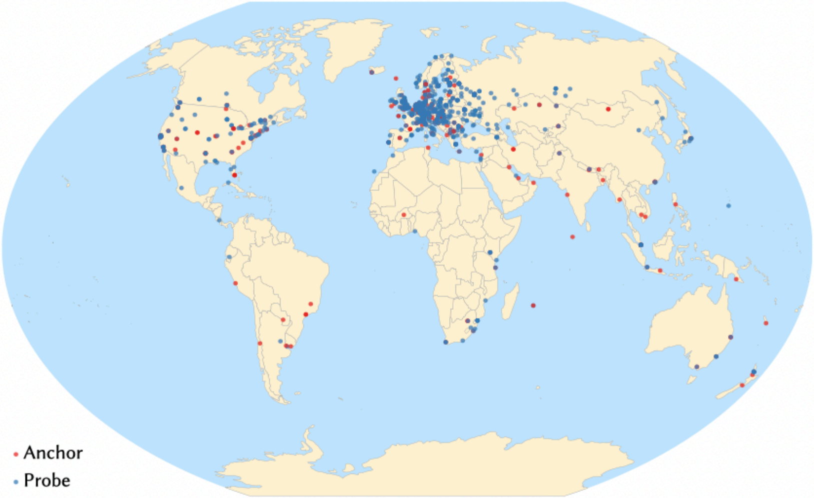
https://atlas.ripe.net



Global population density
as of 2015

GPWv4, CIESIN/SEDAC
http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev10/maps

Testing active geolocation around the world



RIPE Atlas anchors
and stable probes

Crowdsourced test hosts
(40 volunteer, 150 MTurk)

Measurement approach: *crowdsourcing*

- We crowd-sourced hosts in known locations from around the world. *(We make the known “target” probe the landmark, still RTT)*
- 2 campaigns:
  - a) To validate that the tool works properly
    - Linux: command line vs browser1 vs browser2: no notable differences
    - Windows: noisier
  - b) To test and improve the algorithms
    - 40 volunteers + 150 contributors
- User shares their location, runs measurements against landmarks, uploads measurements
- Campaign “b” goal: find an algorithm that would always include the true location (at the expense of returning a broader region)

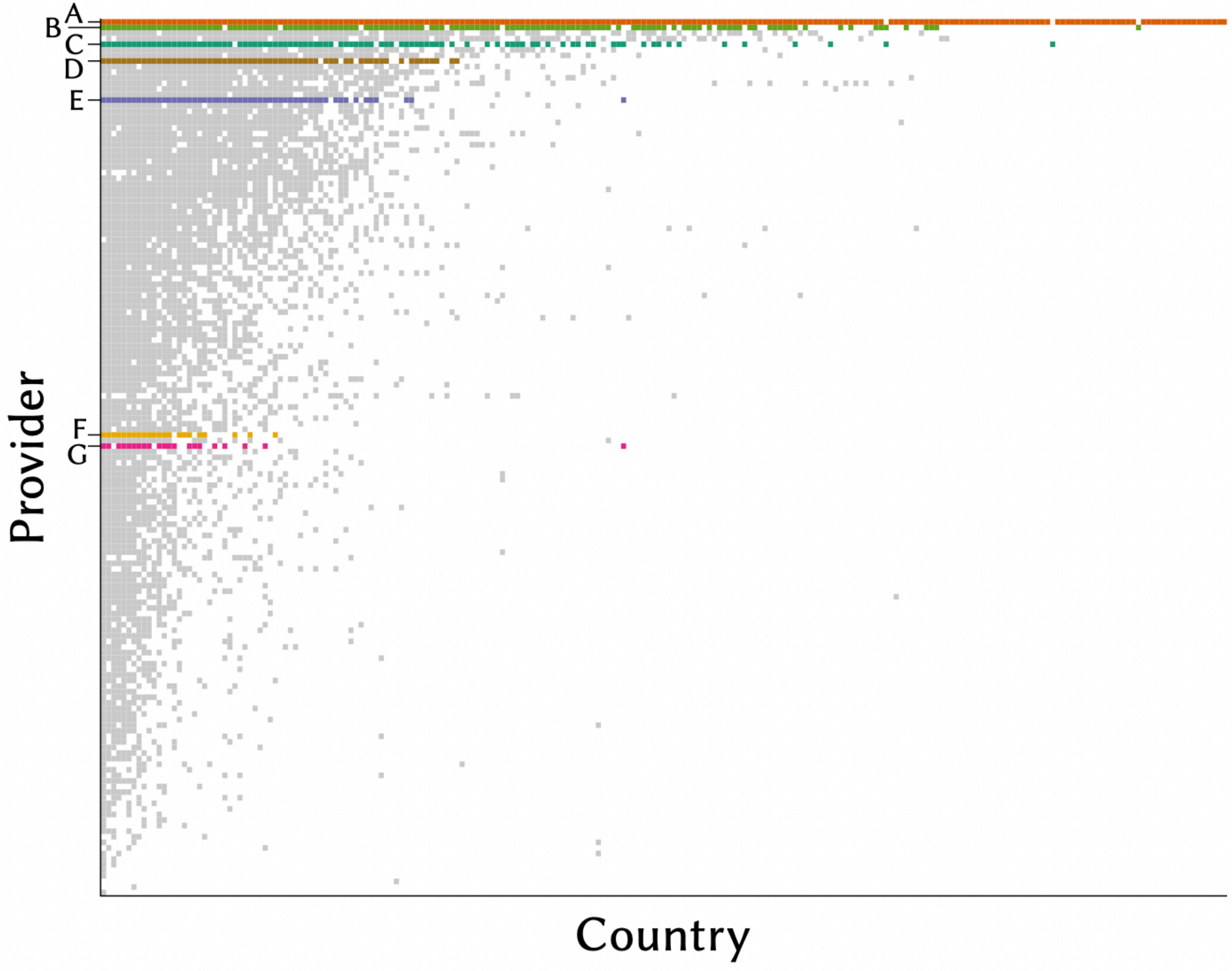
Measurement approach: *CBG++ (Sec. 5.1)*

- CBG is the most effective
  - but doesn't always cover the true location in its prediction
  - Can only fail because disks are too small: underestimates the distance the pkt can travel
    - Can happen due to congestion during calibration.
- Make modifications
  - Travel speed estimates no faster than undersea cable speeds (200km/ms)
  - No slower than 84.5 km/ms <- no landmark can be farther than half the equatorial circumference of Earth
  - More sophisticated multilateration

Experimental analysis

- Tested 7 VPN providers

Seven VPN providers



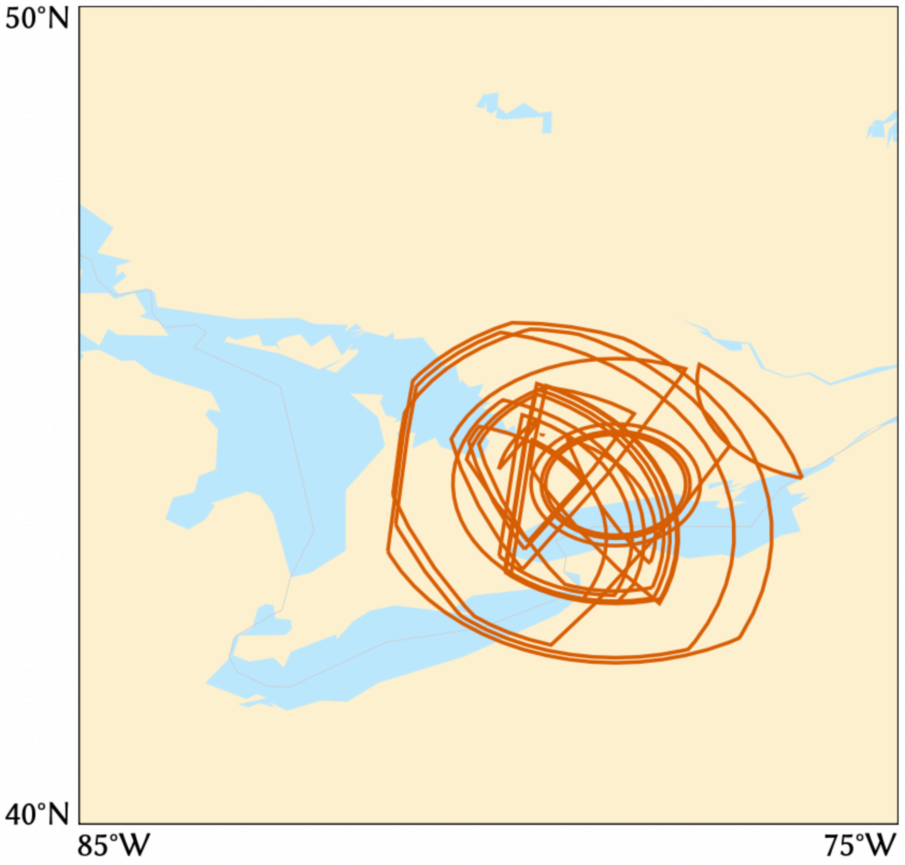
157 VPN providers

VPN commercial landscape data collected by [VPN.com](VPN.com)

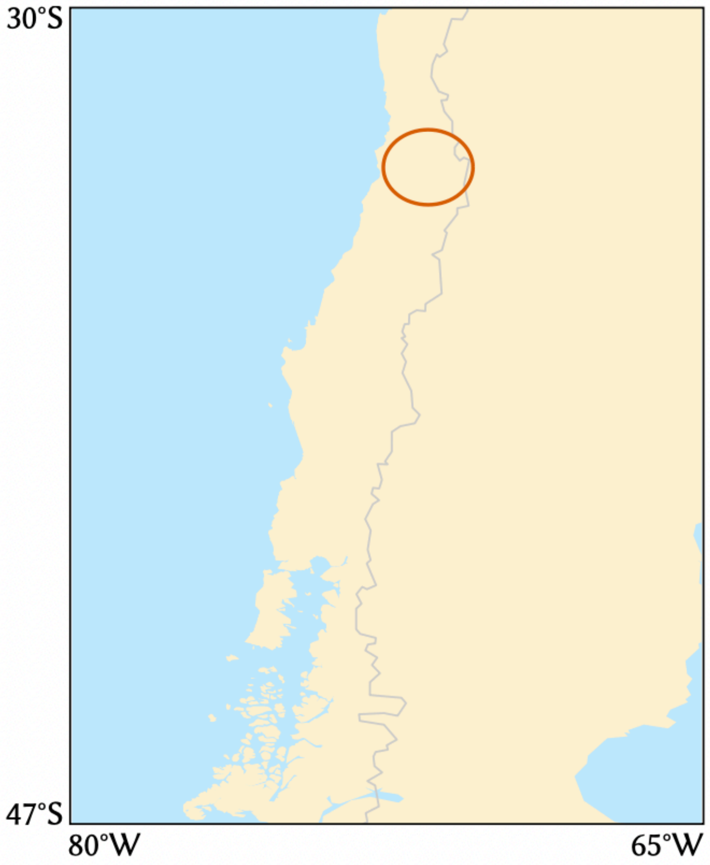
Experimental analysis

- Tested 7 VPN providers
- Tested 2269 unique server IP addresses, allegedly distributed over 222 countries and territories
  - None of the providers advertise exact locations for their proxies
  - We only evaluate **country-level** claims
- Outcomes
  - ***False***: the predicted region does not cover any part of the claimed country
  - ***Credible***: the predicted region is entirely within the claimed country
  - ***Uncertain***: the predicted region covers both the claimed country and others
- Additional criteria used (Sec. 6):
  - Locations of datacenters (leverages the Internet Atlas project)
  - Proxies sharing the same /24 and same origin AS are assumed being in the same datacenter (-> country)

Disambiguation with external knowledge



All these targets belong to the same AS and /24



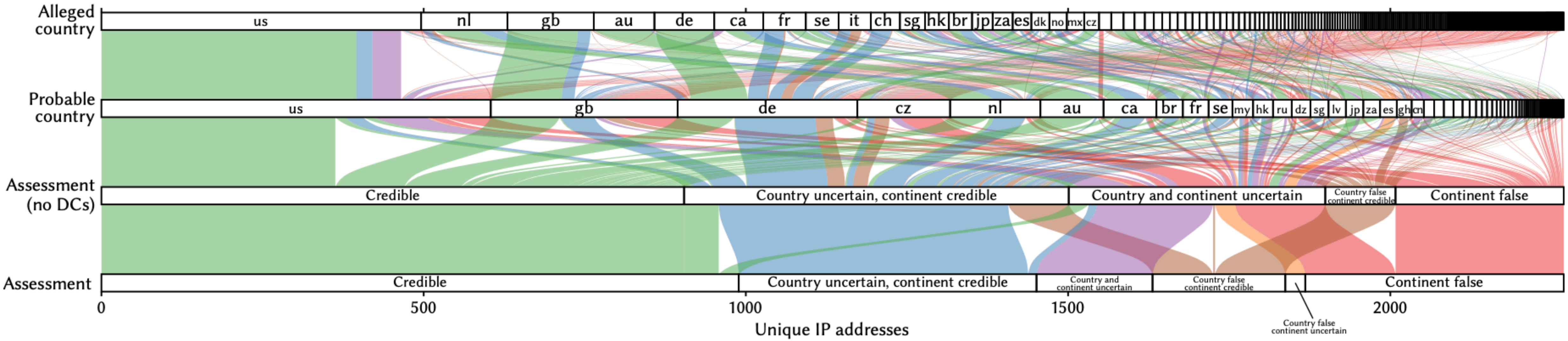
All the data centers inside the oval are in Chile

Experimental results

- Claimed locations (out of 2269 IP addresses)
  - credible:989
  - uncertain: 642
  - false: 638
- For 401 of the false addresses, the true location is not even on the same continent as the claimed location

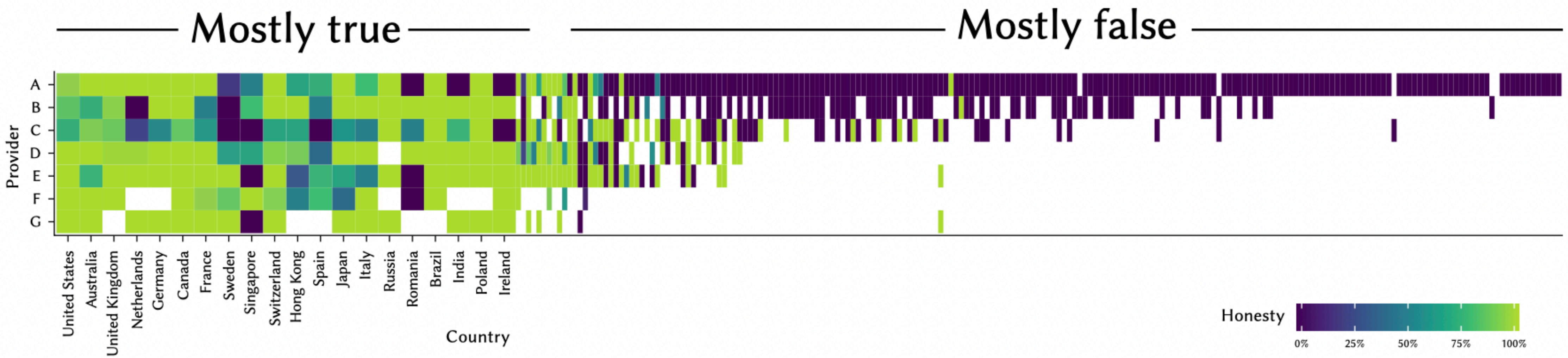
Experimental results

- Which countries are more likely to host credibly-advertised proxies, and where the servers for the false claims actually are
- The 10 countries with the largest number of claimed proxies account for 84% of the credible cases, and only 11% of the false cases.
- False claims are spread over the “long tail” of countries, with only a few advertised servers each.



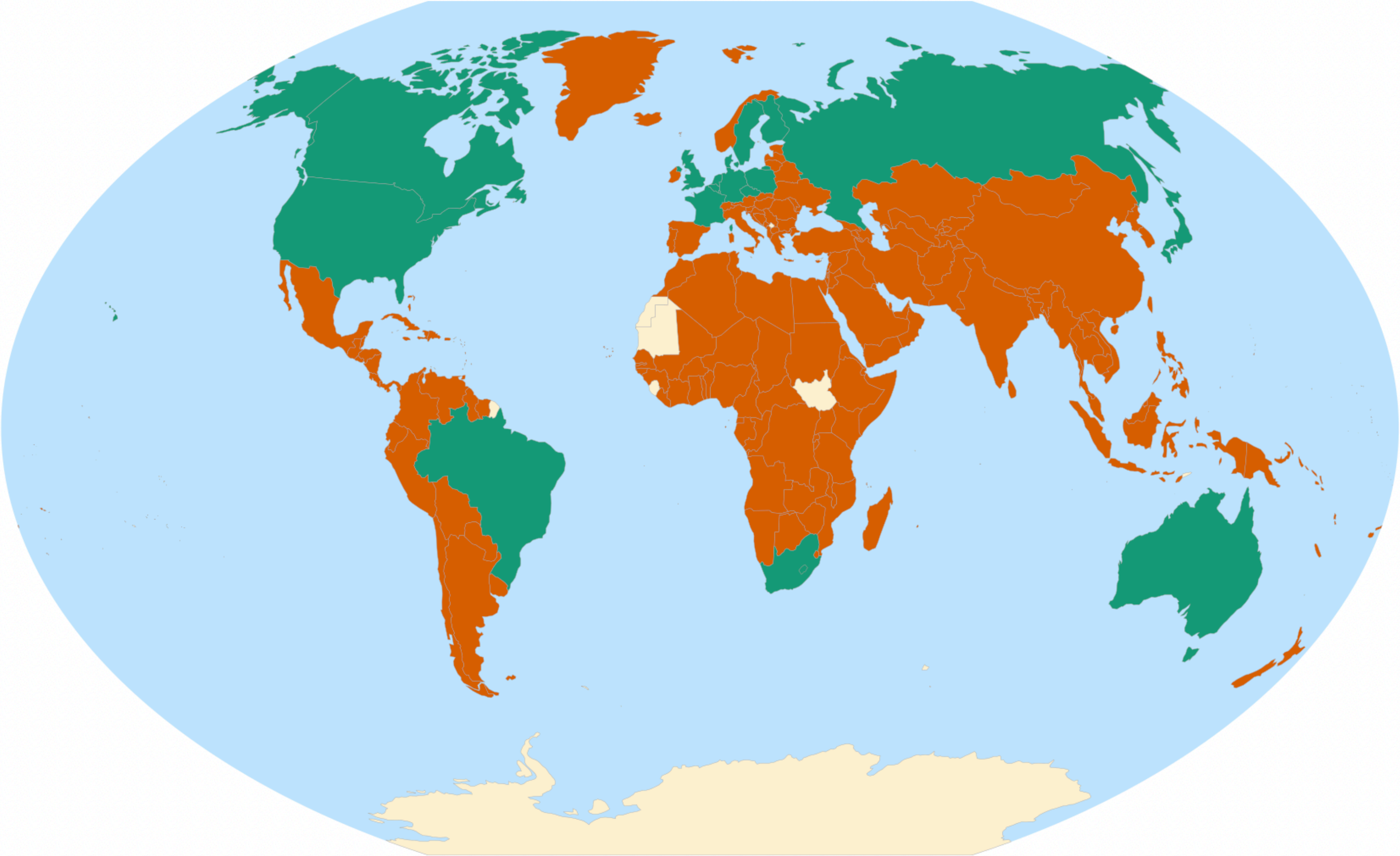
Experimental results

- The credible claims are concentrated in the countries where many other VPN providers also claim to host proxies.
- This is evidence for our original intuition that proxies are likely to be hosted in countries where server hosting is easy to acquire.



Dishonest claims are more likely to occur in the "long tail" of countries.

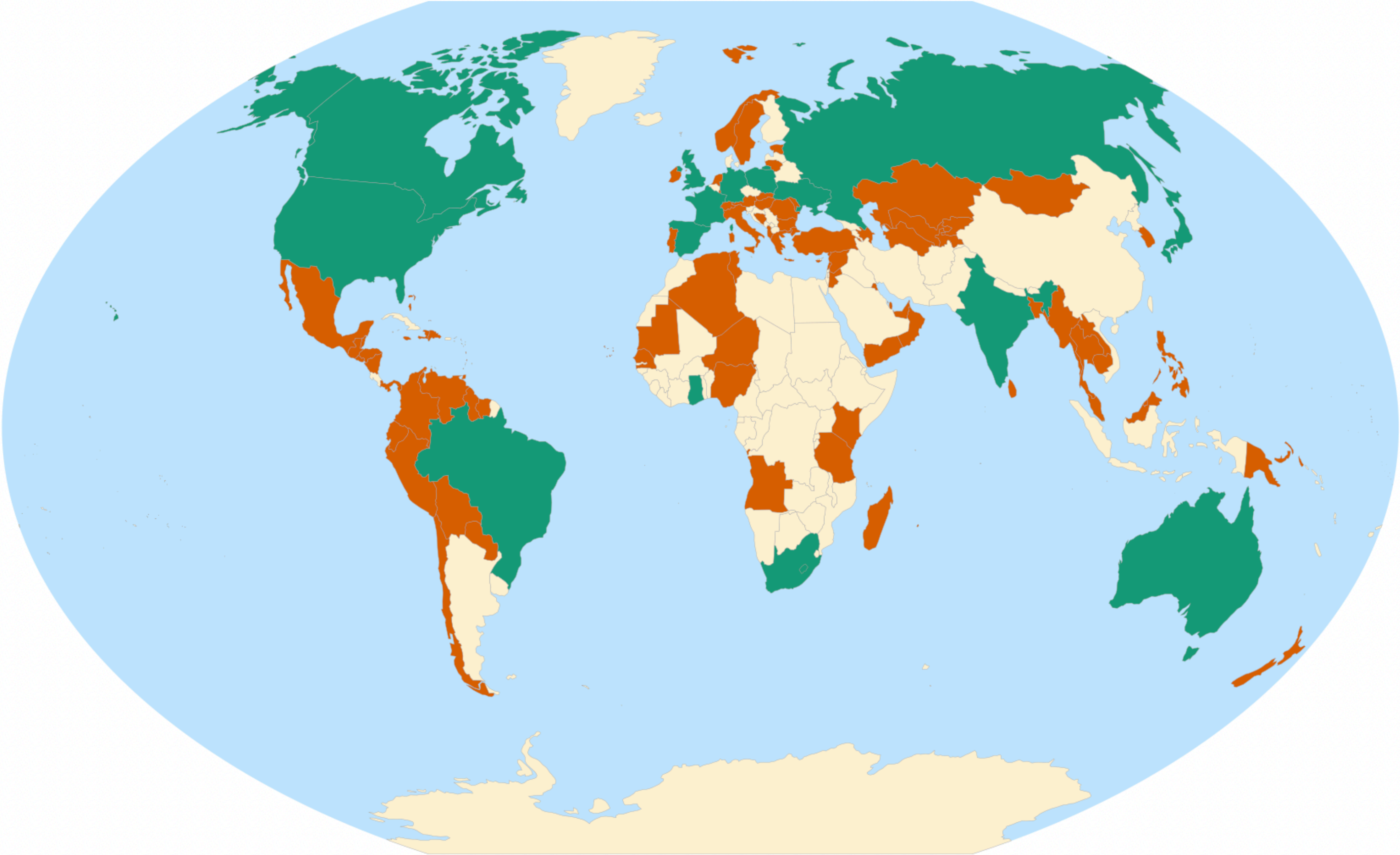
Provider A



Claim 218 countries

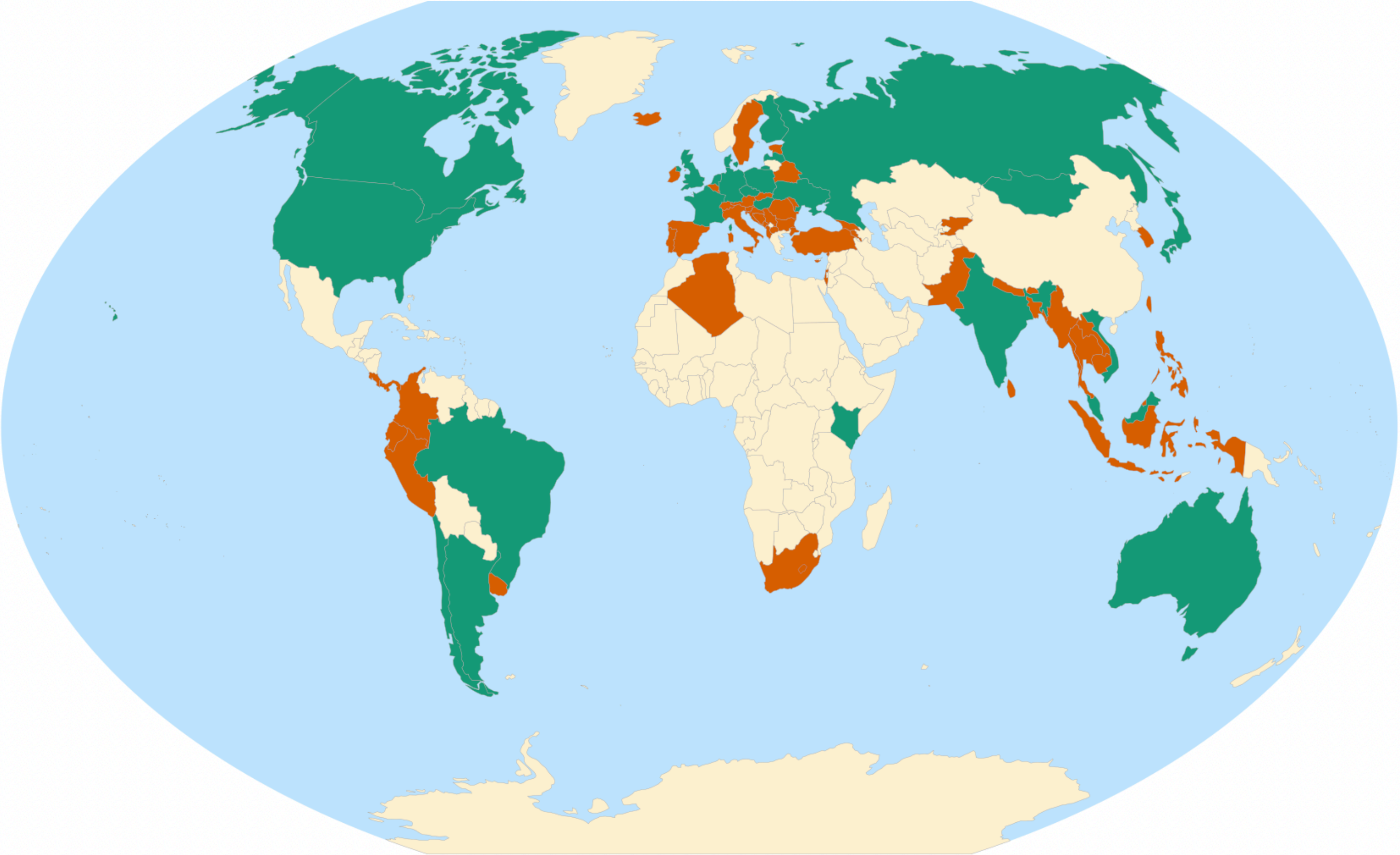
No more than 40 true countries

Provider B



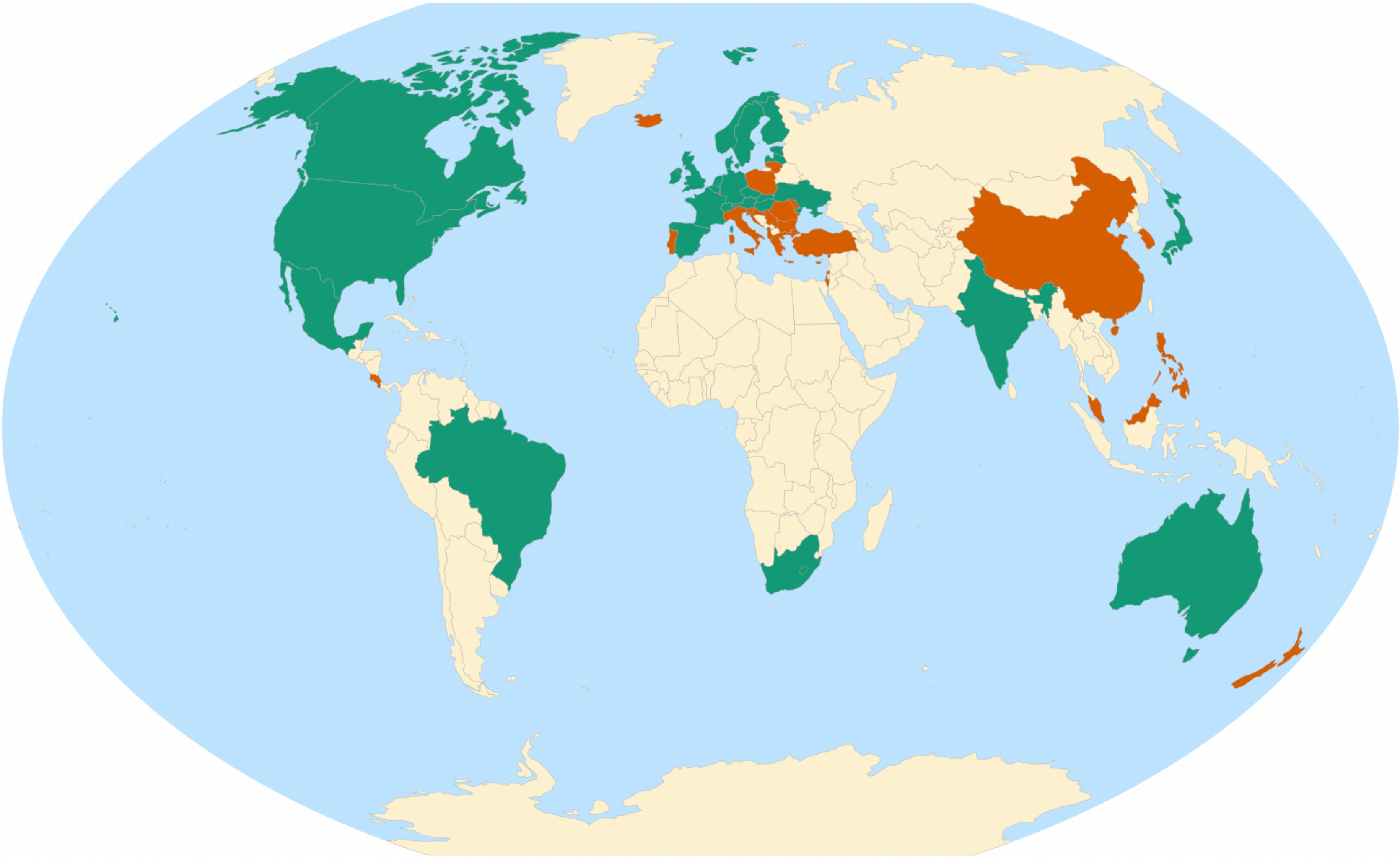
Claim 109 countries
No more than 30 true countries

Provider C



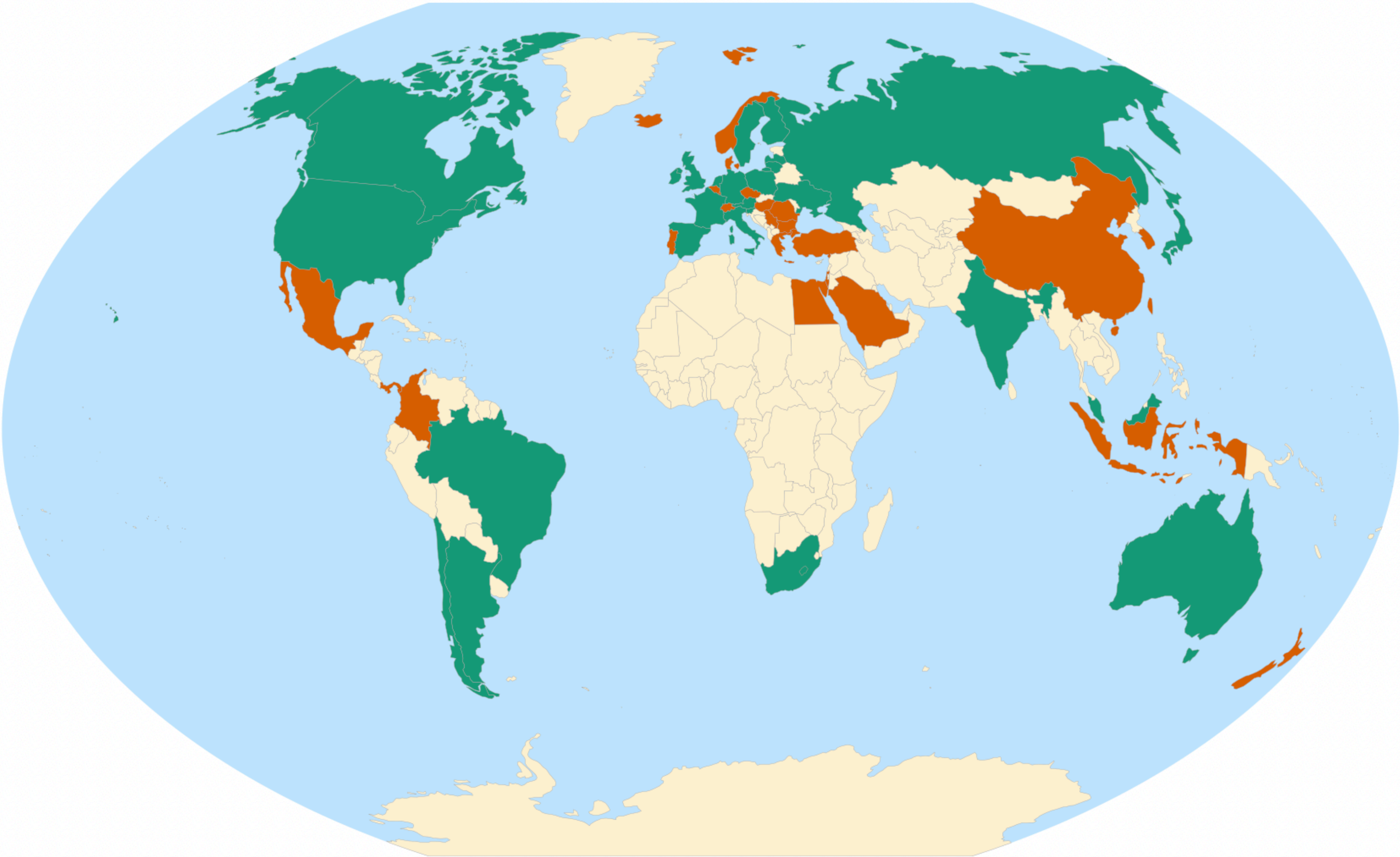
Claim 84 countries
No more than 50 true countries

Provider D



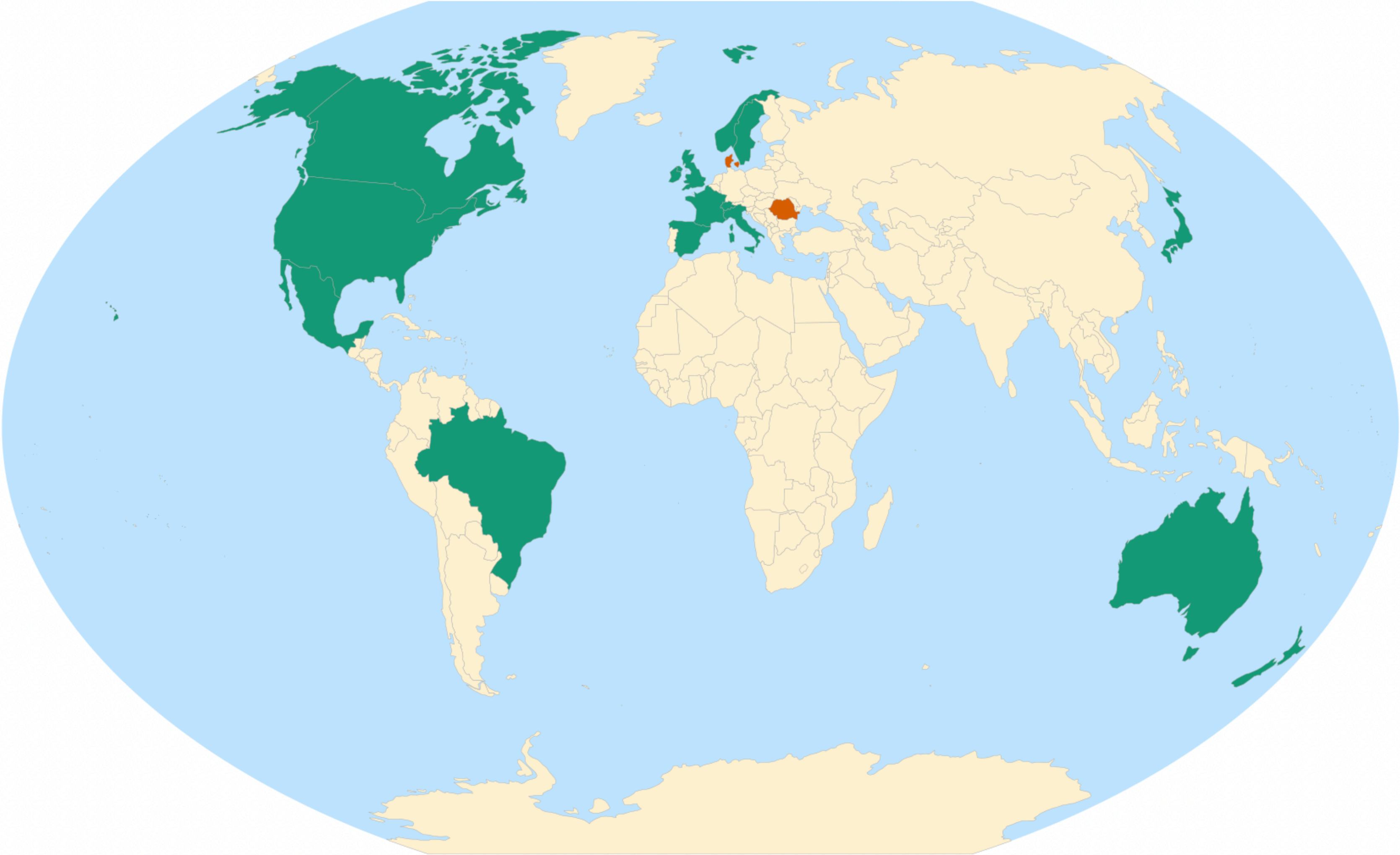
Claim 52 countries
No more than 45 true countries

Provider E



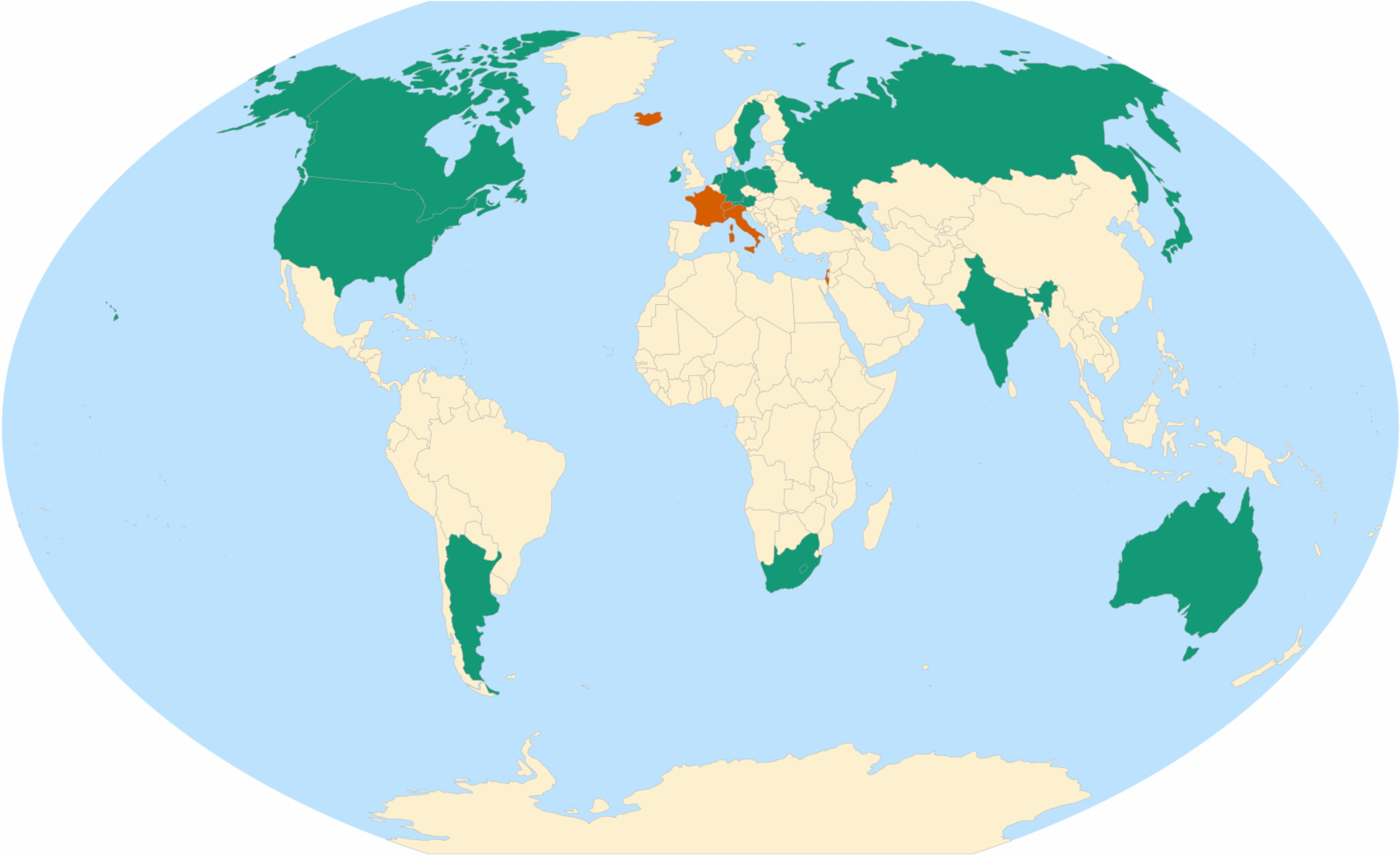
Claim 53 countries
No more than 35 true countries

Provider F



Claim 19 countries
Could be as many as 31 countries

Provider G



Claim 20 countries
No more than 18 true countries

Experimental results

- There is some variation among the providers; for instance, C and E are actually hosting servers in more than one country of South America, whereas providers A and B just say they are.
- However, claimed locations in countries where server hosting is difficult are almost always false.
- Even in regions like Western Europe, where hosting is available in any country one would like, providers seem to prefer to concentrate their hosts in a few locations.

Either we're wrong or the databases are

generous: we assume that all of the "uncertain" cases are actually credible

strict:we assume they are all false

	Provider						
	A	B	C	D	E	F	G
DB-IP	99%	86%	94%	88%	98%	97%	94%
Eureka	99%	99%	99%	82%	99%	100%	100%
IP2Location	47%	65%	91%	77%	95%	97%	91%
IPInfo	39%	93%	97%	79%	97%	93%	100%
MaxMind	99%	99%	99%	82%	99%	100%	100%
CBG (generous)	42%	48%	61%	94%	86%	82%	91%
CBG (strict)	27%	30%	40%	62%	49%	32%	64%

Experimental results

- All five of the IP-to-location databases are more likely to agree with the providers' claims than the active-geolocation approach is.
- We are inclined to suspect that this is because the proxy providers have influenced the information in these databases. We have no hard evidence backing this suspicion
  - but we observe that there is *no pattern* to the countries for which the IP-to-location databases disagree with provider claims. This is what we would expect to see if the databases were being influenced, but with some lag-time.
- As the proxy providers add servers, the databases default their locations to a guess based on IP address registry information, which, for commercial data centers, may be reasonably close to the truth. When the database services attempt to make a more precise assessment, this draws on the source that the providers can influence.

Discussion (from the actual paper's section)

- All providers declined to respond
- Results call into question validity of measurements that leverage VPNs to gain location diversity
- Many customers might be content to appear in a country independently from the actual truth
- Deliberate false information?
- Potential interference with RTT measurements
- Web-based measurement technique could be used to geolocate visitors without their knowledge :-(
- Future work
  - more providers (there are >150 !)
  - Trying to make the Web-based tool as accurate as the command-line one

What I liked

- Code is available!
  - *https://github.com/zackw/active-geolocator*
- Topic with very practical implications
  - brings up questions about policies, risks for the users, …
- Another useful finding that impacts IDS: that we can't blindly trust VPNs for certain measurements!
- Re. the class:
  - brings up the topic of crowdsourcing measurements
  - some interesting viz
- I wish they dug into the “lies”. E.g., checking whois data from registries and domain names

(Example of) What did I learn

- Another type/source of errors in geolocation DBs
- Measurements from VPN servers might be affected by severe errors
- Several sources of landmarks
  - RIPE Atlas locations are skewed
- Can use crowdsource measurements
  - Crowdsourced measurements bring several challenges
    - Less controlled experiment
    - …
- A method that can be applied to verify geolocation in general?
- Implications for privacy of web users

Thanks

Datasets/tools

- *"Using the 2012 Natural Earth map of the world, we also exclude oceans and lakes"*
- RIPE Atlas
- Internet Atlas
- Geolocation DBs: *Maxmind, …*
- Mechanical Turk
- *…?*