Creation of the Data Infrastructure Dataset

Marco Giovanni Ferrari Prasiddha Rajaure

2025-03-28

Abstract

This code produces the dataset collecting informations on regional (NUTS2) data infrastructure. We merge different sources regarding regional statistics, electric infrastructure, datacenters, landing points and bathymetric data.

Introduction

The following code performs the creation of the map of the world data network. In particular it considers cables and cloud datacenters available in the maps provided by TeleGeography[1]. It aggregates observations at country level in a network where:

1. The ***nodes*** are single countries or territorial area. We define as unique territorial areas as those areas that, conditional being under the same authority, are within the distance of 100Km to one another. In the commented code in `code/world_boundaries_definition.Rmd` we detail how we compute these areas that are then stored in `work_data/world_data_infrastructure/world_boundaries_single_territories.Rmd`.
2. The ***edges*** are the subsea cables connecting the pair of nodes. The *thickness* of each edge represent the number of cables connecting them. We consider two countries as connected if there exist a cable connecting them *without passing through other countries*. We then sum the number of direct available paths to obtain the measure of the thickness, this allow us to evidence which countries are the so called *choke points* that hold the highest number of connection with the rest of the world. We compute the number of nodes connecting countries in the commented code in `code/world_data_infrastructure/distance_matrix_extraction.Rmd` whose result are stored in `work_data/world_data_infrastructure/distance_dataset.Rds`.
3. The ***size (radius) of the nodes*** can be either:
   - The number of total connections that each node has;
   - The number of data centers present withing that geographical boundary. We consider all the three types of cloud datacenters provided in the TeleGeography Cloud Infrastructure Map[2]. They are namely:
     - **Cloud Regions**: groups of data centers that are the primary units in the functioning of the global cloud industry.
     - **Local Zones**: data centers built with the purpose of serving more remote geographical areas with low latency.
     - **On Ramps**: data centers that have the purpose to connect the local (and terrestrial) data infrastructure with the global (submarine) one.

[1] For subsea cables: https://www.submarinecablemap.com/
For cloud data centers: https://www.cloudinfrastructuremap.com/
[2] Source: https://www.cloudinfrastructuremap.com/

The data on datacenters location are scraped from the webpage of TeleGeography in the code `code/datacenter_location_extraction.Rmd` and stored in the three files, corresponding to each type of datacenter, available in the folder `work_data/datacenters_location`.

We plot this network in a world map. In doing so, we consider the spherical shape of earth, in order to accurately reproduce the actual distance between nodes.

Importing data

In this preliminary part we import the data and show how they are initially structure, we proceed also to some preliminary manipulation.

```
edges_raw <- readRDS("work_data/world_data_infrastructure/distance_dataset.rds")

edges_raw <- edges_raw %>%
  rowwise() %>%
  mutate(
    node1 = min(c(A, B)),
    node2 = max(c(A, B))
  ) %>%
  ungroup()
```

Edges then appear as follow:

Table 1: Edges Dataset

A	B	cable	distance	node1	node2
SGP_1	THA_1	sea-h2x-0	1129134.9	SGP_1	THA_1
PHL_1	SGP_1	sea-h2x-0	2847129.0	PHL_1	SGP_1
PHL_1	THA_1	sea-h2x-0	3057199.3	PHL_1	THA_1
MYS_1	SGP_1	sea-h2x-0	1285652.8	MYS_1	SGP_1
MYS_1	PHL_1	sea-h2x-0	2650147.4	MYS_1	PHL_1
MYS_1	THA_1	sea-h2x-0	1495723.0	MYS_1	THA_1
MYS_1	MYS_2	sea-h2x-0	1135889.3	MYS_1	MYS_2
MYS_2	SGP_1	sea-h2x-0	339224.5	MYS_2	SGP_1
MYS_2	PHL_1	sea-h2x-0	2697365.5	MYS_2	PHL_1
MYS_2	THA_1	sea-h2x-0	979371.4	MYS_2	THA_1

We also have the information on the year of Ready-For-Service of each cable, which we merge to the edge dataset. We can then subset our dataset and see the evolution across time of the global data network.

```
cable_info <- read_excel("raw_data/subsea_cables_location/cable_info.xlsx") %>%
  drop_na(year)

month_to_quarter <- function(x) {
  x_lower <- str_to_lower(x)
  case_when(
    str_detect(x_lower, "jan|feb|mar") ~ "Q1",
    str_detect(x_lower, "apr|may|jun") ~ "Q2",
    str_detect(x_lower, "jul|aug|sep") ~ "Q3",
    str_detect(x_lower, "oct|nov|dec") ~ "Q4",
    str_detect(x_lower, "q[1-4]") ~ str_to_upper(x),
```

```
    # When no info on the month is available, we automatically assign Q4
    TRUE ~ "Q4"
  )
}

cable_info <- cable_info %>%
  mutate(
    quarter = month_to_quarter(month),
    quarter = factor(quarter, levels = c("Q1", "Q2", "Q3", "Q4"), ordered = TRUE)
  ) %>%
  relocate(quarter, .after = year) %>%
  rename(cable = feature_id) %>%
  select(cable, year, quarter)

edges_raw <- edges_raw %>%
  left_join(cable_info, by = join_by(cable)) %>%
  drop_na(year)
```

Since our dataset, which is dated 2022, contains also future and under construction cables, we restrict the dataset to the existing ones (i.e., those ready for service before 2022).

```
edges <- edges_raw %>% filter(year < 2022)
```

For this moment we ignore the effective length of cables and we limit our interest to the number of available direct connections:

```
edges <- edges %>%
  count(node1, node2, name = "total_cables")
```

Which can be organized in the form of adjacency matrix:

```
edges_network <- graph_from_data_frame(edges, directed = FALSE)

edges_adj_matrix <- as.data.frame(as_adjacency_matrix(edges_network,
                                                      attr = "total_cables",
                                                      sparse = FALSE))
```

Table 2: Adjacency Matrix (first 10 rows and columns)

	ABW_1	AGO_1	ALB_1	ARE_1	ARG_1	ASM_1	ATG_1	AUS_2	AUS_4	BEL_1
ABW_1	0	0	0	0	0	0	0	0	0	0
AGO_1	0	0	0	0	0	0	0	0	0	0
ALB_1	0	0	0	0	0	0	0	0	0	0
ARE_1	0	0	0	0	0	0	0	0	1	0
ARG_1	0	0	0	0	0	0	0	0	0	0
ASM_1	0	0	0	0	0	0	0	0	1	0
ATG_1	0	0	0	0	0	0	0	0	0	0
AUS_2	0	0	0	0	0	0	0	0	1	0
AUS_4	0	0	0	1	0	1	0	1	0	0
BEL_1	0	0	0	0	0	0	0	0	0	0

We import then the position of the territories and extract the centroid to compute the map:

```
nodes <- readRDS("work_data/world_data_infrastructure/world_boundaries_single_territories.rds")

world_boundaries_single_territories <- nodes

nodes <- nodes %>%
  st_centroid() %>%
  mutate(x = st_coordinates(.)[, 1], y = st_coordinates(.)[, 2]) %>%
  st_set_geometry(NULL) %>%
  relocate(country_code_n, .after = y ) %>%
  select(-country_code)
```

Then we assign to each node the number of total connections:

```
nodes_size <- edges_adj_matrix %>%
  mutate(total_connections = rowSums(.)) %>%
  select(total_connections) %>%
  rownames_to_column(var = "country_code_n") %>%
  left_join(nodes, by = join_by(country_code_n))
```

Table 3: Node Size

country_code_n	total_connections	x	y
ABW_1	5	-69.98	12.52
AGO_1	23	17.54	-12.27
ALB_1	2	20.07	41.13
ARE_1	59	54.33	23.90
ARG_1	6	-64.75	-34.68
ASM_1	4	-170.71	-14.30
ATG_1	2	-61.79	17.28
AUS_2	3	105.70	-10.44
AUS_4	37	134.30	-25.77
BEL_1	4	4.67	50.64

We also import data on datacenters:

```
datacenter_count <- readRDS("work_data/datacenters_location/datacenters_count.rds") %>%
  select(country_code_n, Total_DataCenters) %>%
  rename(total_datacenters = Total_DataCenters)

datacenter_count <- datacenter_count %>%
  left_join(nodes, by = join_by(country_code_n))
```

Merging nodes and edges

In this second part, we proceed by combining the information available in the those datasets.

```
colnames(edges) <- c("from", "to", "total_cables")
colnames(nodes) <- c( "x", "y", "country_code_n")

# Joining edges with node coordinates
```

Table 4: Node Size

country_code_n	total_datacenters	x	y
ARE_1	17	54.33	23.90
ARG_1	5	-64.75	-34.68
AUS_4	67	134.30	-25.77
AUT_1	9	14.12	47.60
BEL_1	8	4.67	50.64
BHR_1	4	50.56	26.02
BRA_1	46	-53.21	-10.62
CAN_1	53	-96.52	60.44
CHE_1	20	8.23	46.80
CHL_1	21	-71.04	-36.18

```
edges_sf <- edges %>%
  left_join(nodes, by = c("from" = "country_code_n")) %>%
  rename(x1 = x, y1 = y) %>%
  left_join(nodes, by = c("to" = "country_code_n")) %>%
  rename(x2 = x, y2 = y)
```

In this way we obtain a dataset that for each connection between nodes expresses:

Table 5: Edges Endpoints

from	to	total_cables	x1	y1	x2	y2
ABW_1	COL_1	1	-69.98	12.52	-73.07	3.89
ABW_1	GBR_1	1	-69.98	12.52	-64.53	18.44
ABW_1	NLD_1	2	-69.98	12.52	-68.69	12.19
ABW_1	PAN_1	1	-69.98	12.52	-80.10	8.51
AGO_1	BEN_1	1	17.54	-12.27	2.34	9.64
AGO_1	BRA_1	1	17.54	-12.27	-53.21	-10.62
AGO_1	CIV_1	2	17.54	-12.27	-5.56	7.63
AGO_1	CMR_1	2	17.54	-12.27	12.74	5.68
AGO_1	COD_1	1	17.54	-12.27	23.65	-2.86
AGO_1	COG_1	1	17.54	-12.27	15.23	-0.84

We define a function to draw the line connecting two nodes based on the coordinates of the two nodes. This function includes the great-circle path, namely it consider the earth as rounded rather then flat, and draws the lines connecting nodes accordingly.

```
drawing_paths <- function(x1, y1, x2, y2, n = 20) {
  # Ensure input is numeric
  x1 <- as.numeric(x1); y1 <- as.numeric(y1)
  x2 <- as.numeric(x2); y2 <- as.numeric(y2)

  # Compute great-circle path
  coords <- gcIntermediate(c(x1, y1), c(x2, y2), n = n, addStartEnd = TRUE, breakAtDateLine = TRUE)

  # If coords is composed of two matrix return a multi-linestring
  # This is the case for those linestrings passing the dateline.

  if (is.matrix(coords)) {
    return(st_linestring(coords))
```

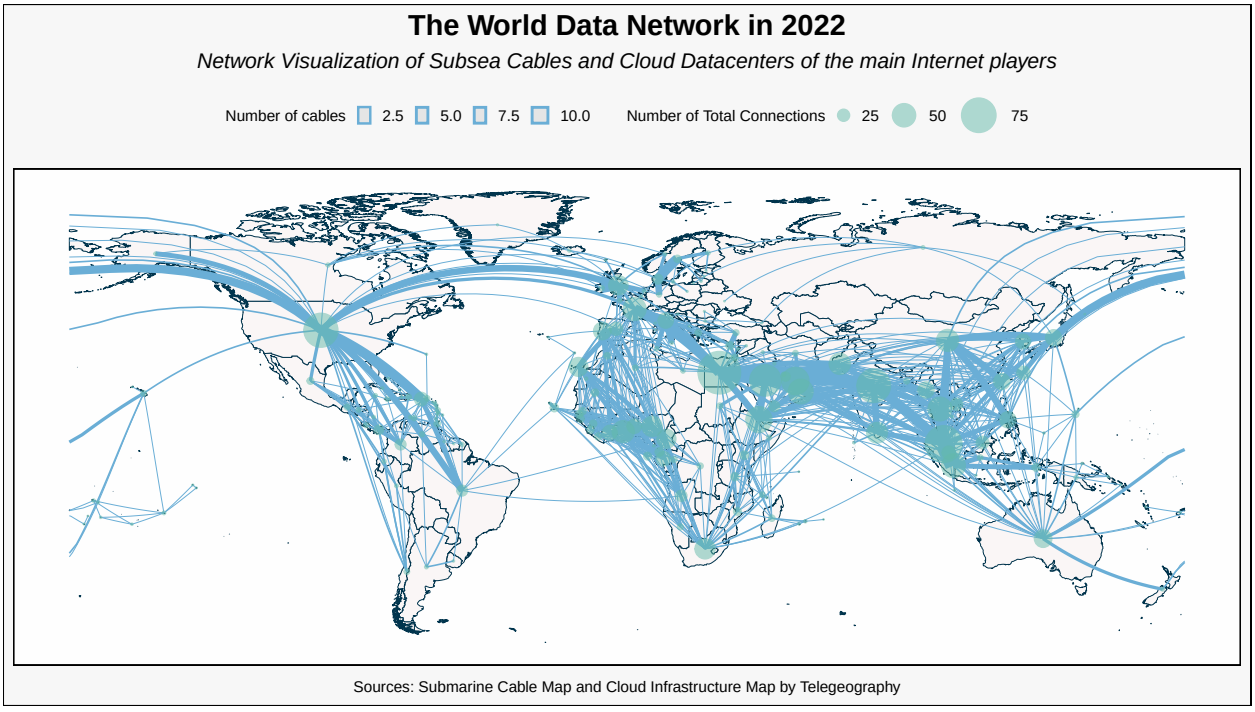
```
  } else if (is.list(coords) && length(coords) == 2) {
    return(st_multilinestring(list(st_linestring(coords[[1]]), st_linestring(coords[[2]]))))
  } else {
    return(st_linestring(matrix(nrow = 0, ncol = 2)))
  }
}
```

Which we appply to our dataframe of endpoints:

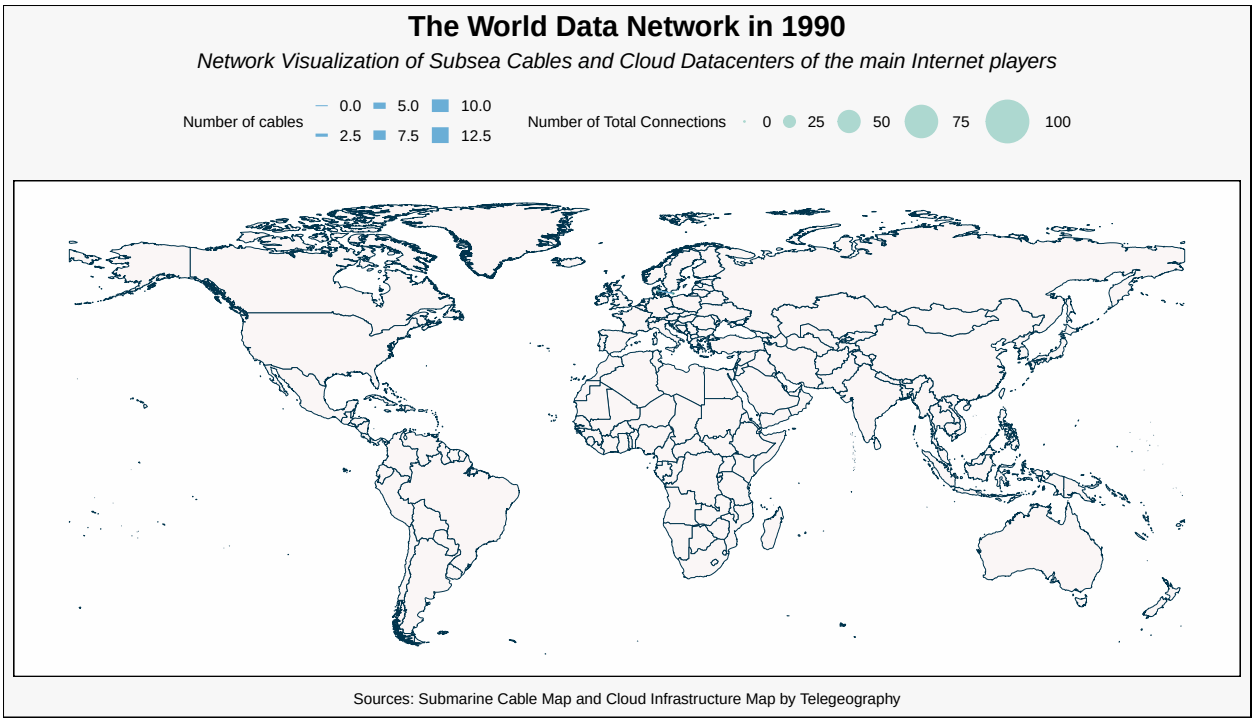
```
edges_sf <- edges_sf %>%
  mutate(geometry = st_sfc(mapply(drawing_paths,
                                  x1, y1, x2, y2,
                                  SIMPLIFY = FALSE),
                           crs = 4326)) %>%
  select(from, to, total_cables, geometry)
```

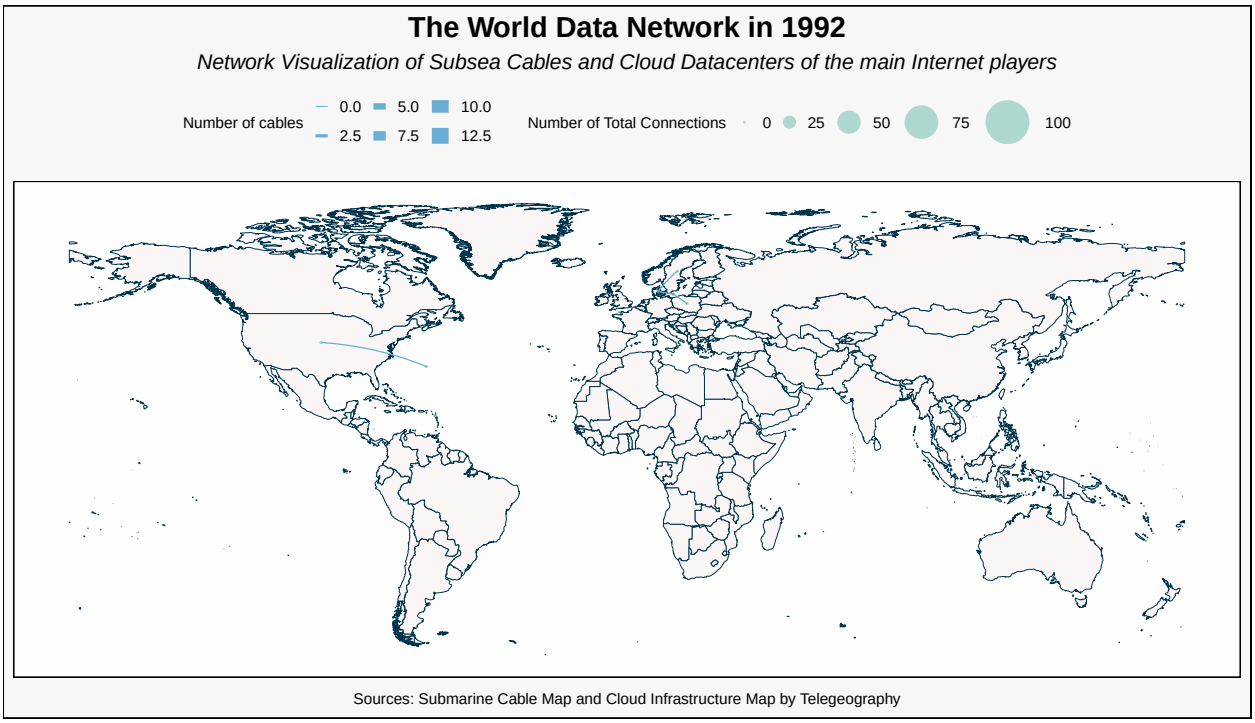
Map creation

We proceed by plotting our for year 2022, and then from the first available data on subsea cables:



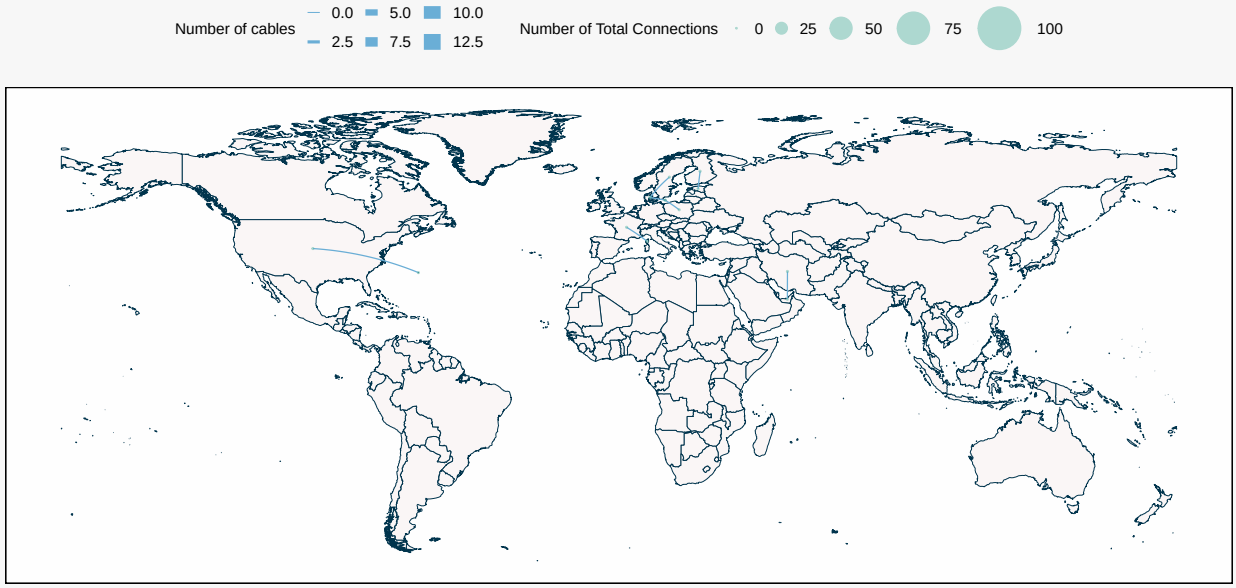
Cable connections in the history





The World Data Network in 1993

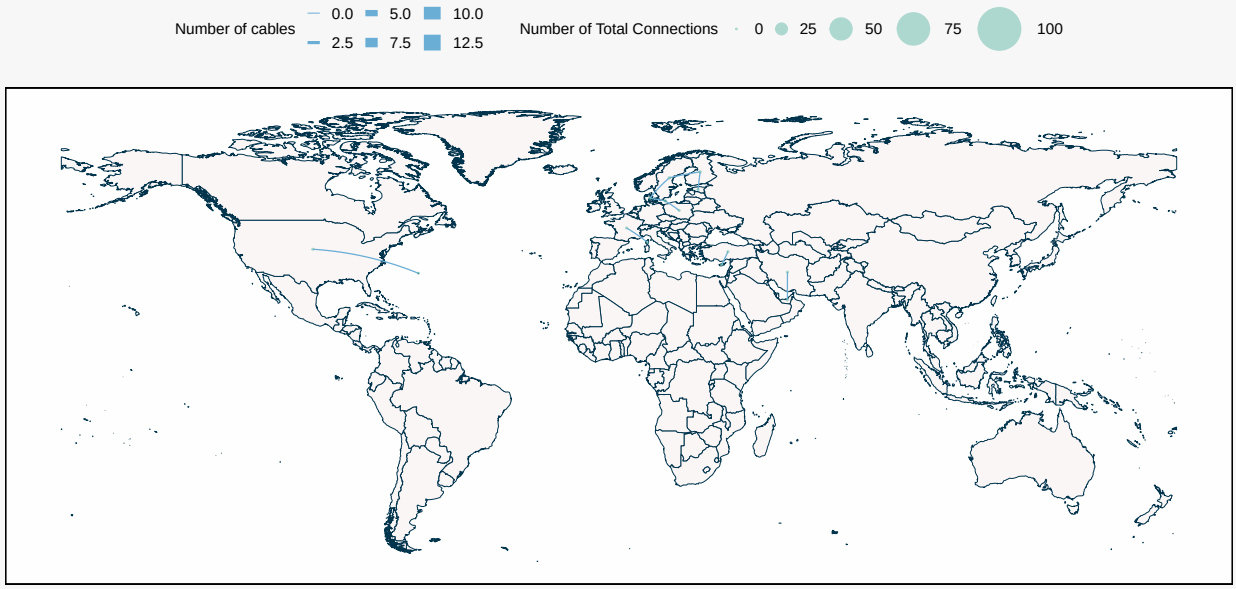
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1994

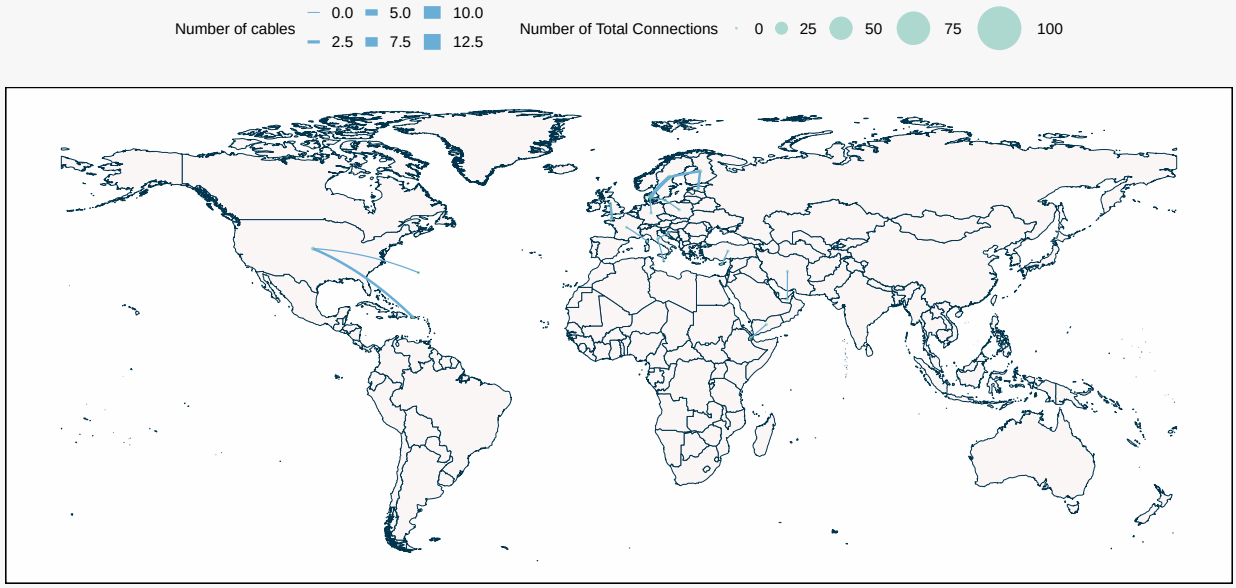
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1995

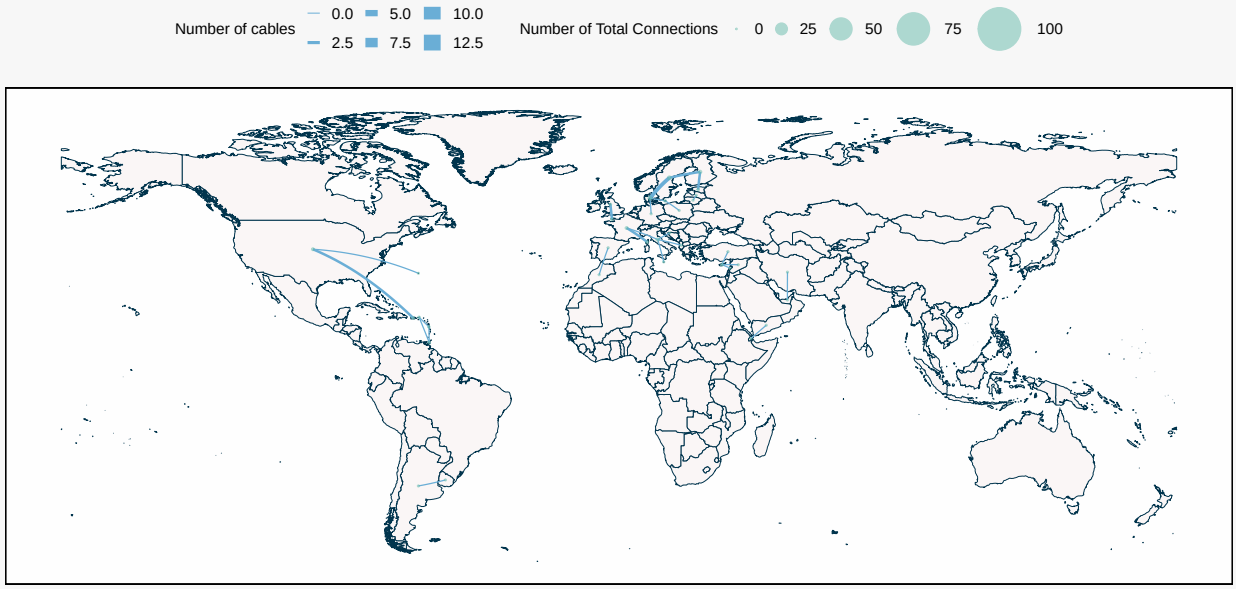
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1996

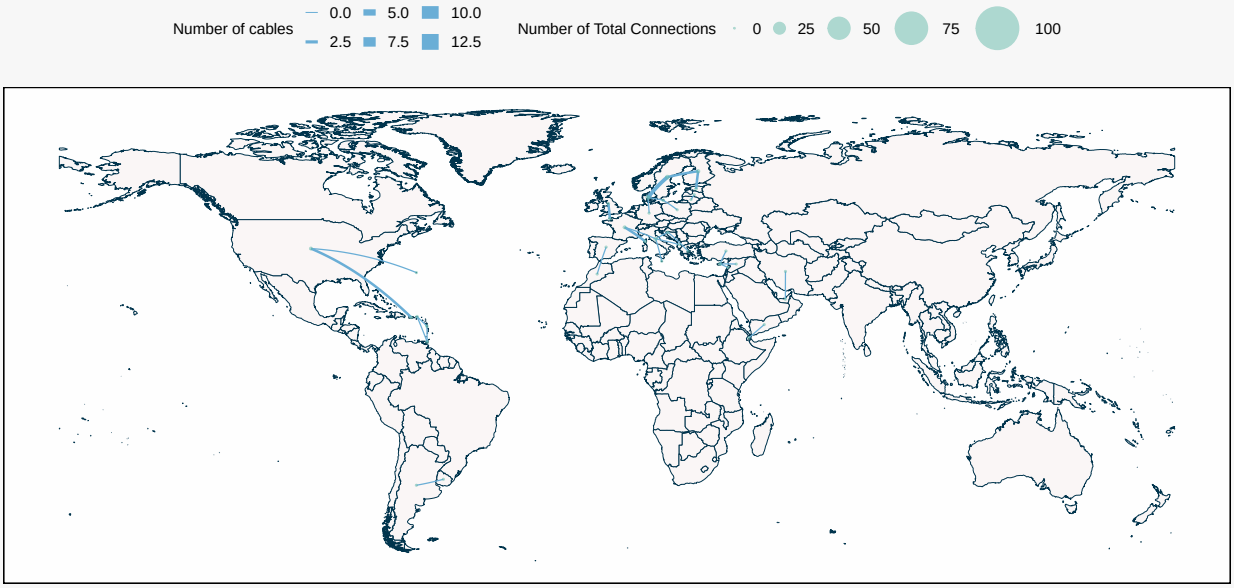
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1997

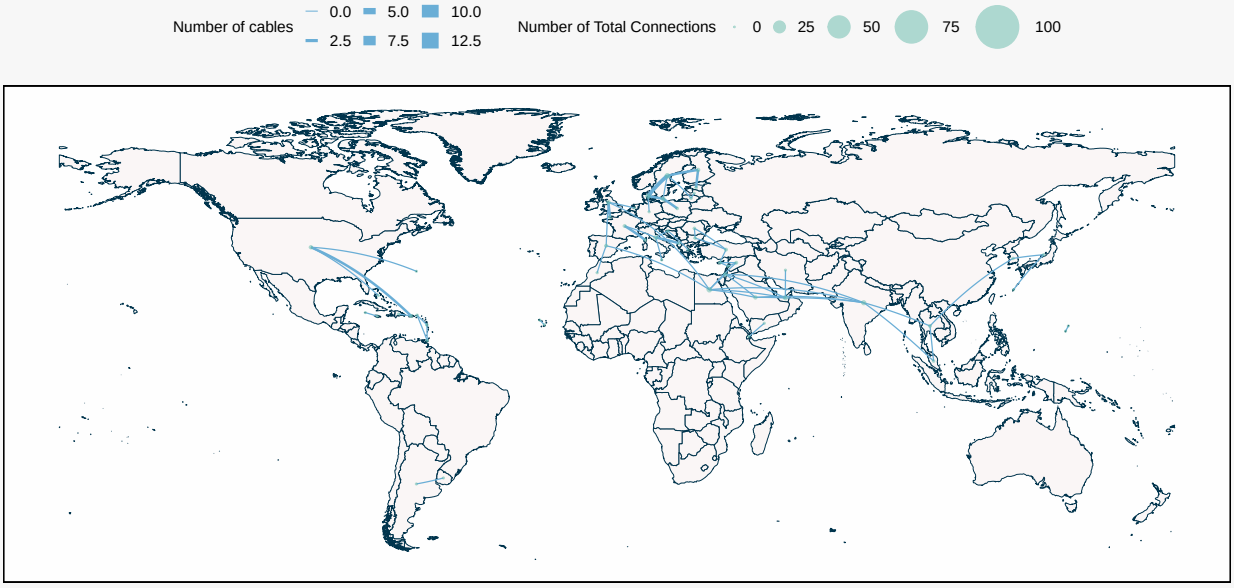
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1998

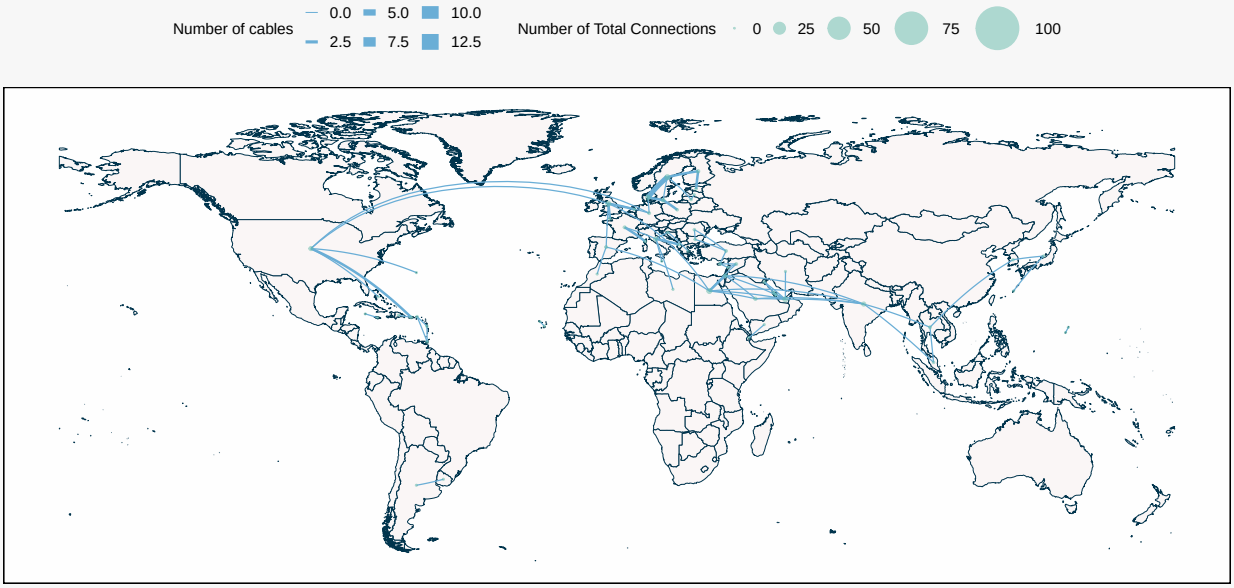
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 1999

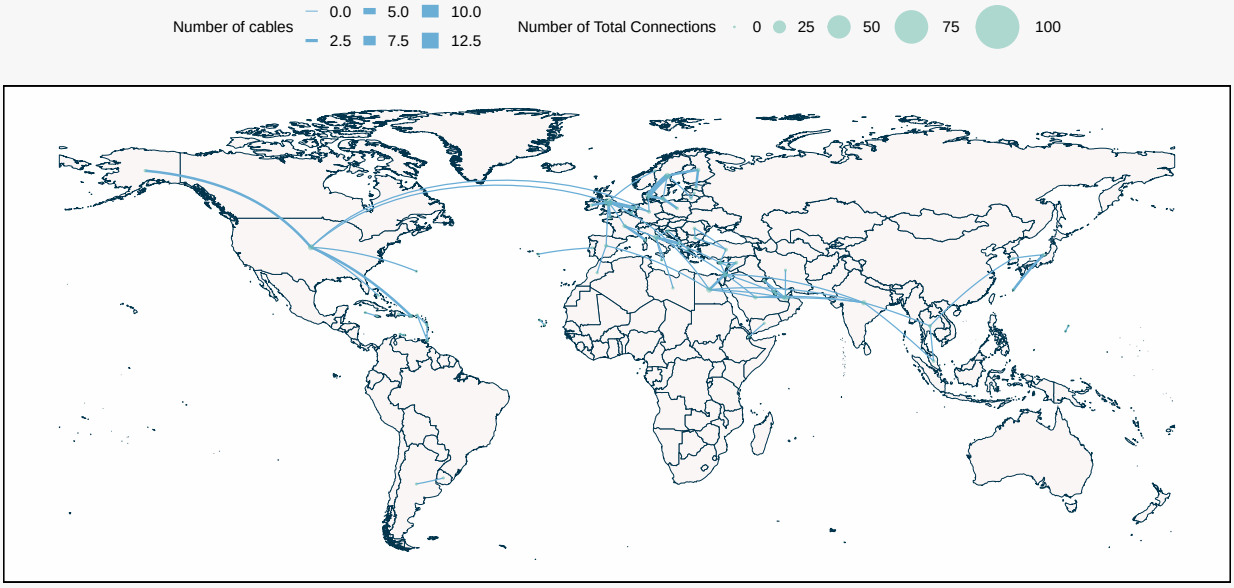
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2000

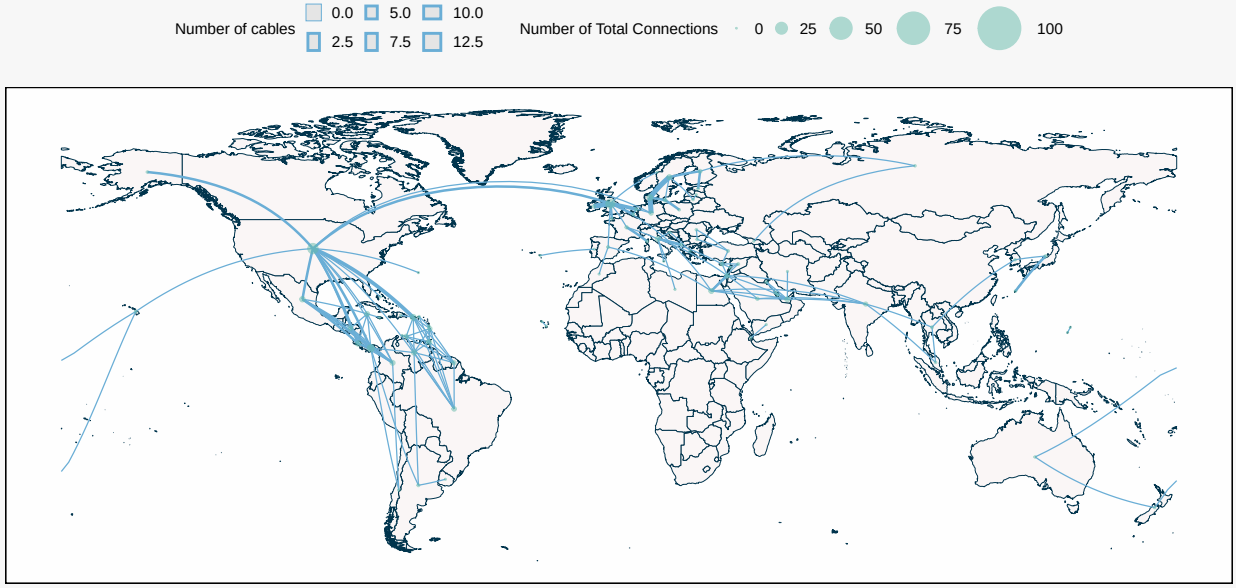
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2001

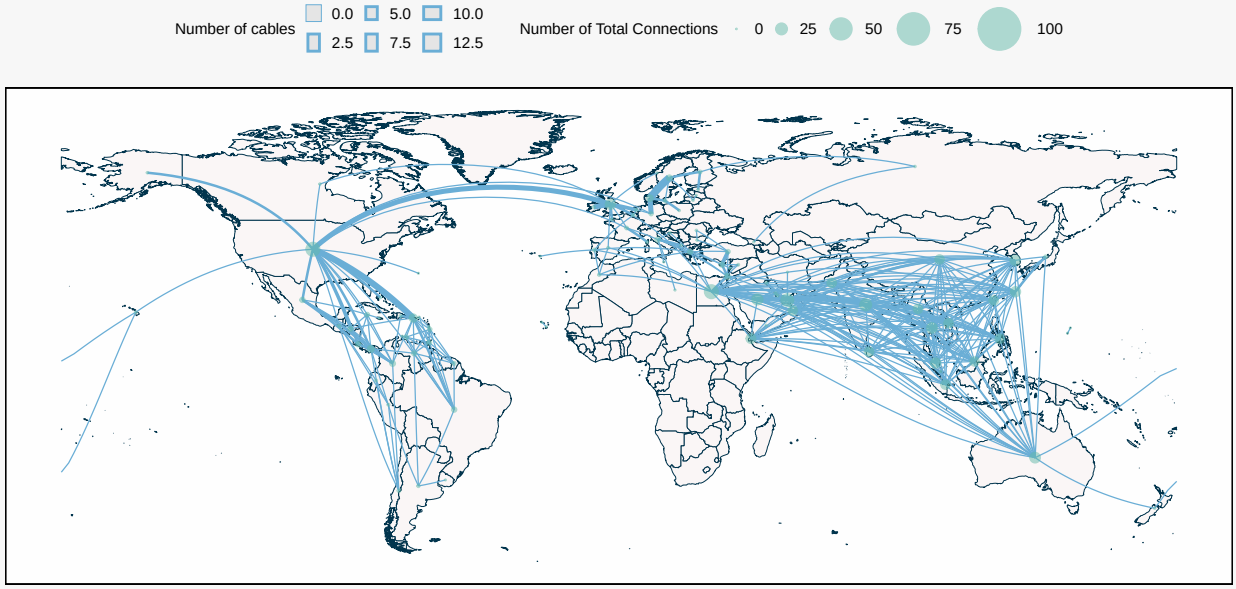
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



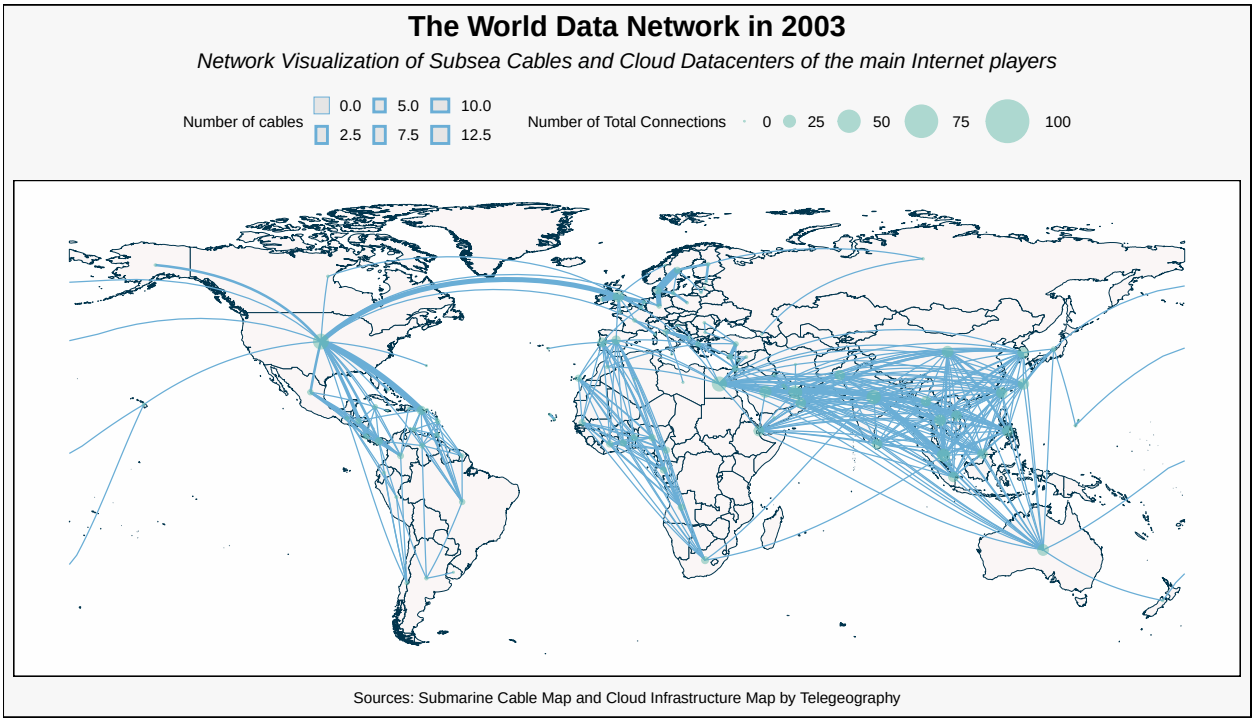
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

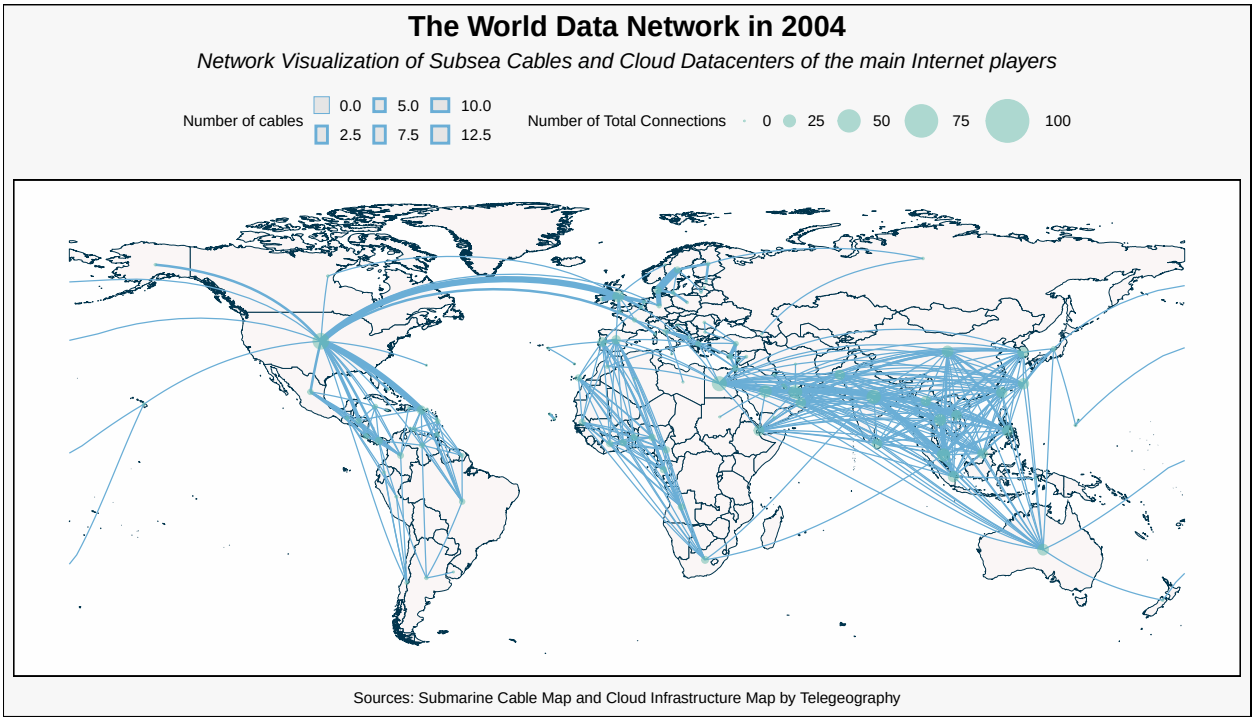
The World Data Network in 2002

Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



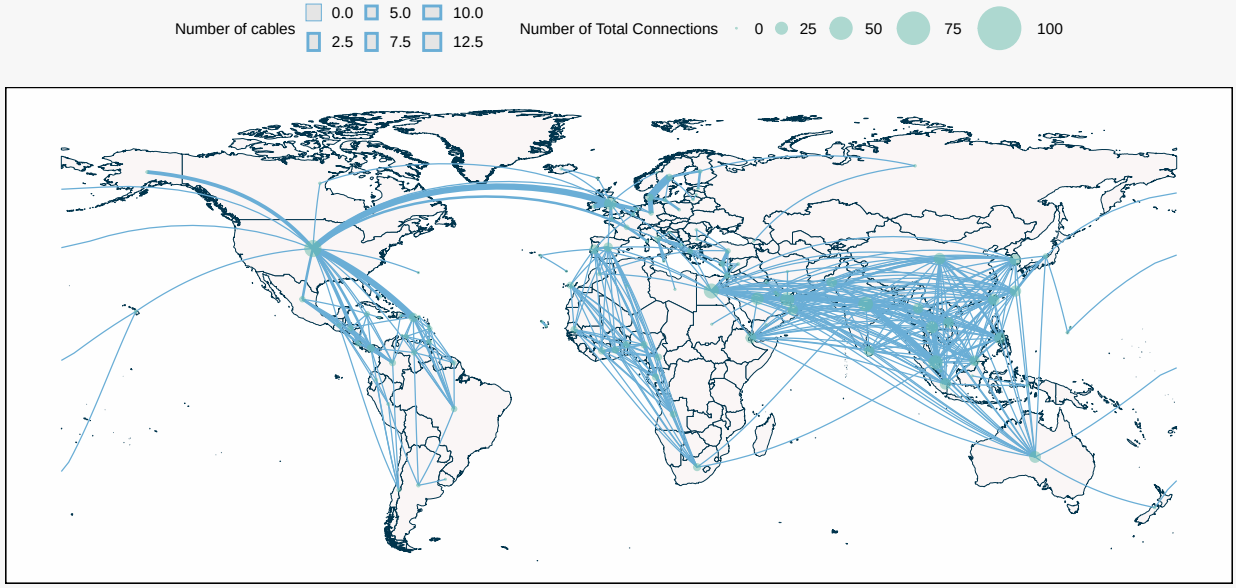
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography


The World Data Network in 2003
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players
Number of cables
0.0
2.5
5.0
7.5
10.0
12.5
Number of Total Connections
0
25
50
75
100
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography


The World Data Network in 2004
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players
Number of cables
0.0
2.5
5.0
7.5
10.0
12.5
Number of Total Connections
0
25
50
75
100
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2005

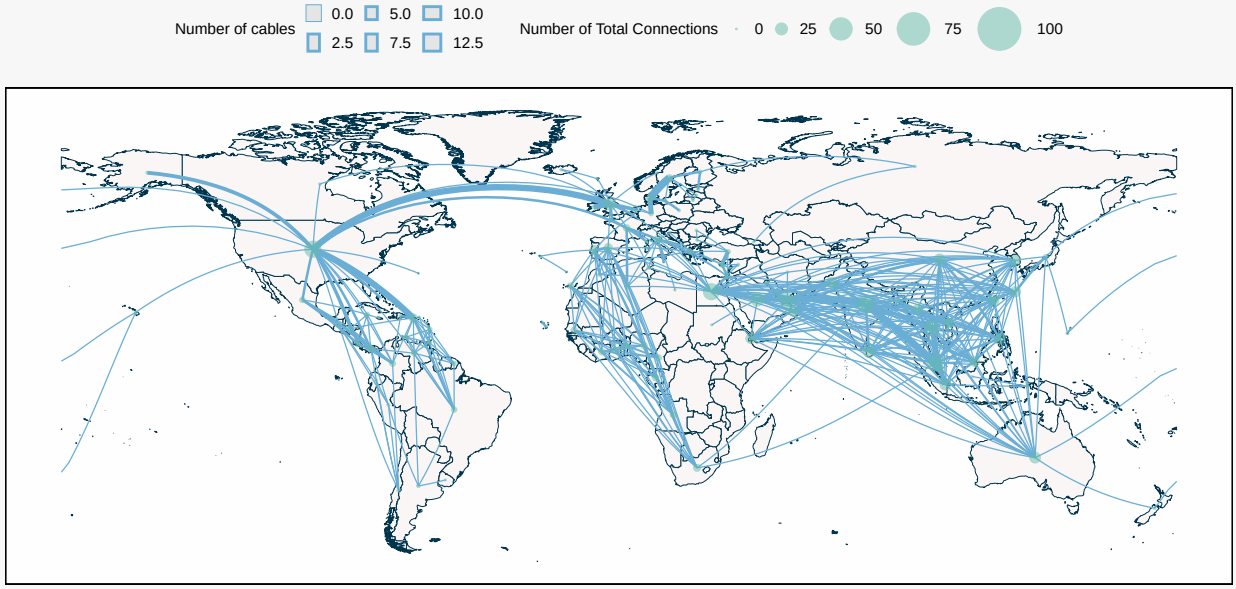
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



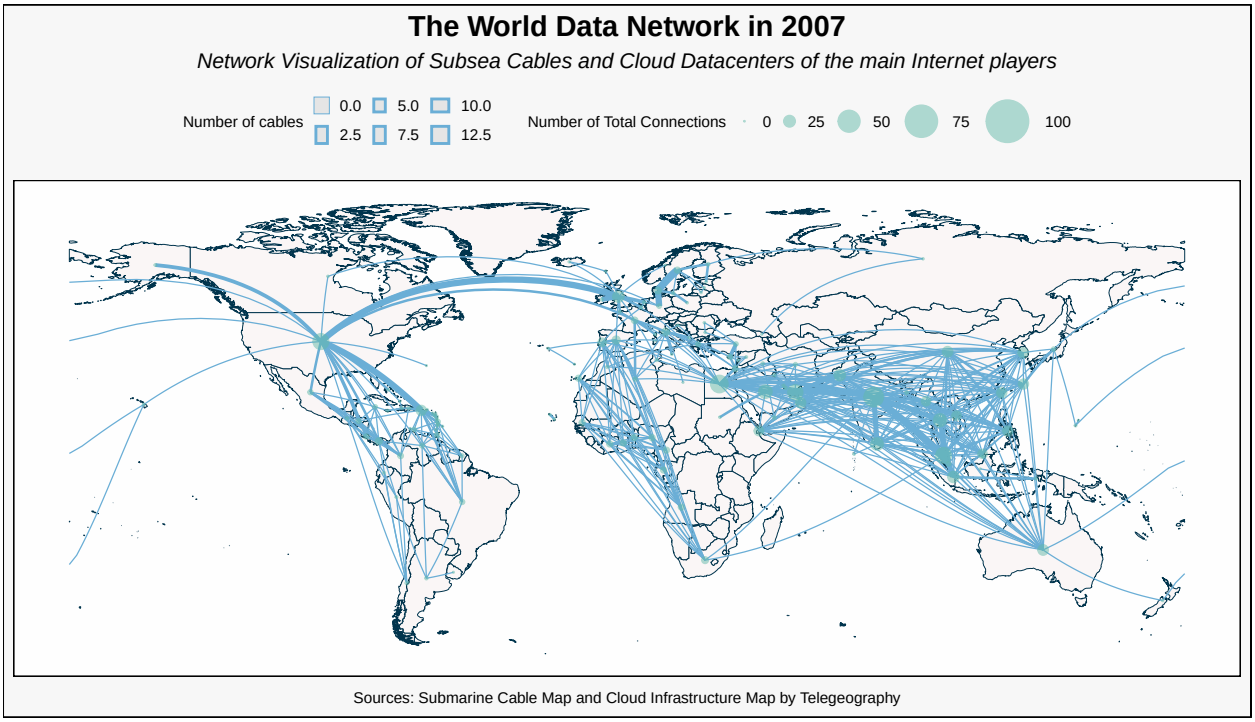
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

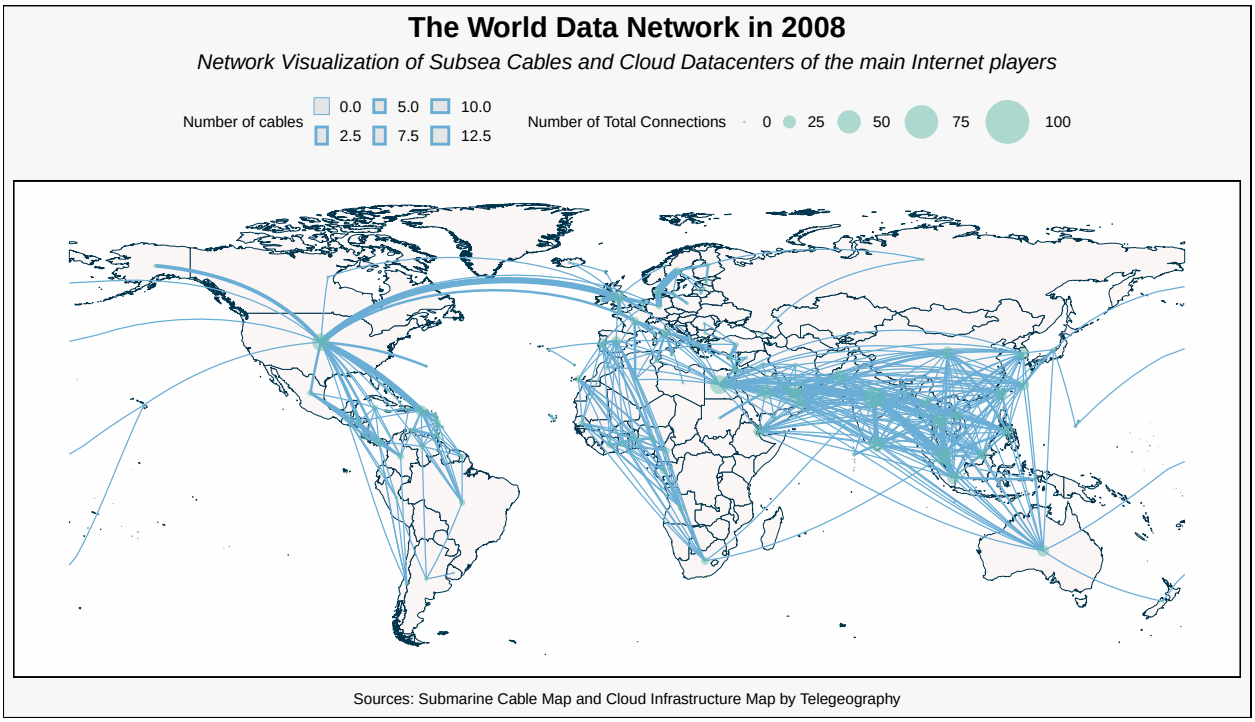
The World Data Network in 2006

Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



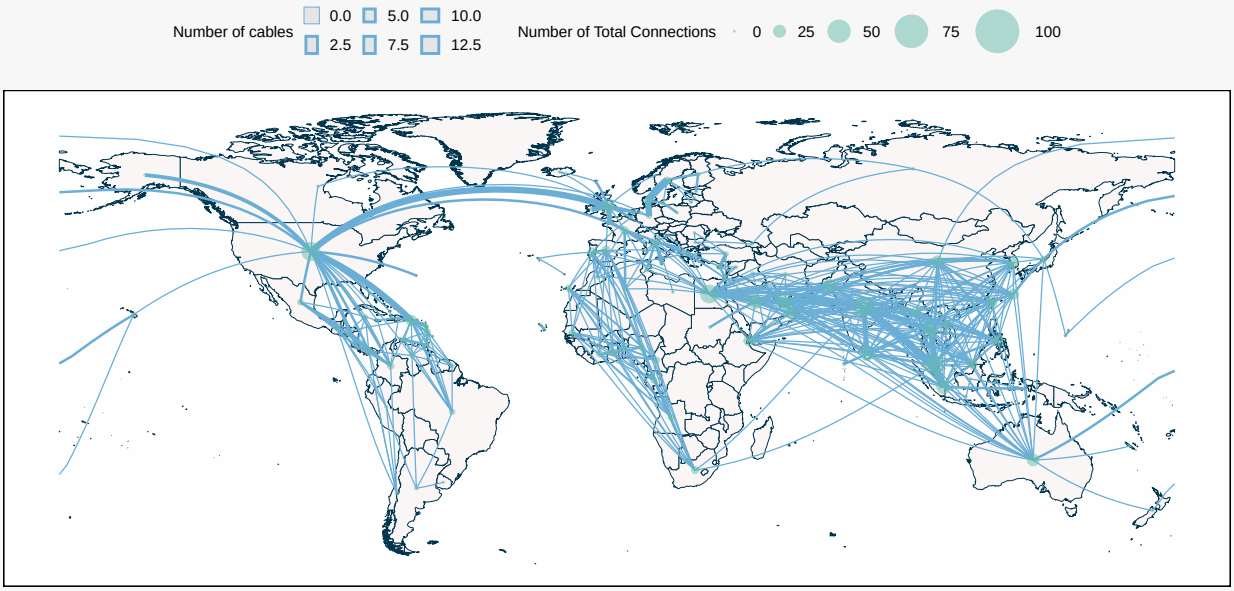
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography


The World Data Network in 2007
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players
Number of cables
0.0
2.5
5.0
7.5
10.0
12.5
Number of Total Connections
0
25
50
75
100
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography


The World Data Network in 2008
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players
Number of cables
0.0
2.5
5.0
7.5
10.0
12.5
Number of Total Connections
0
25
50
75
100
Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2009

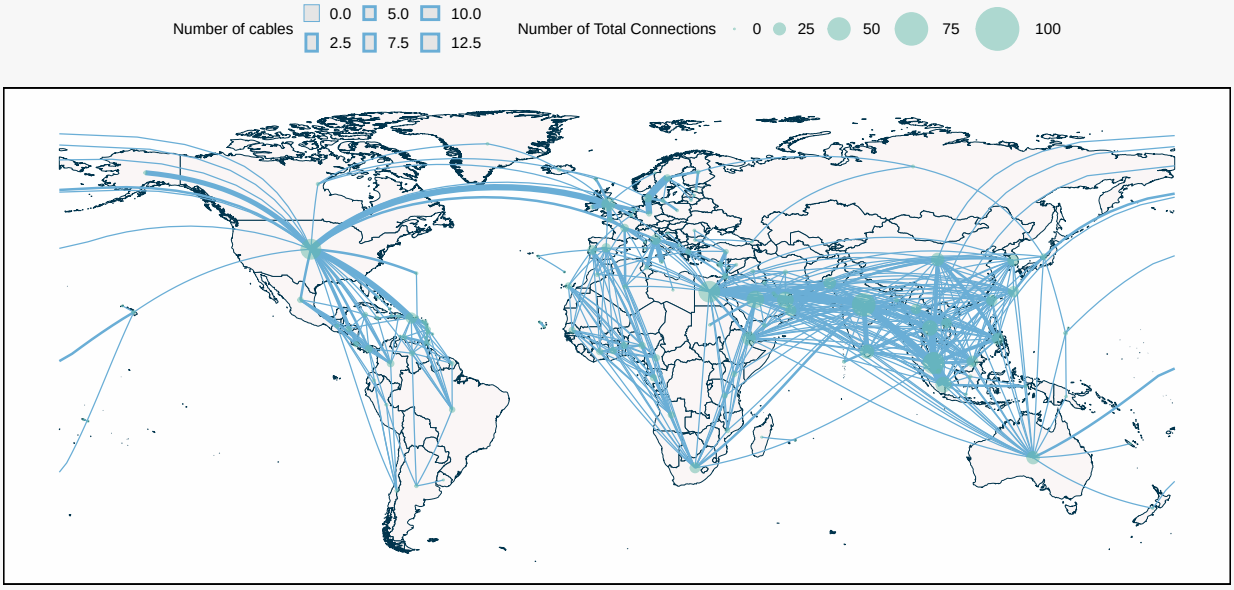
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2010

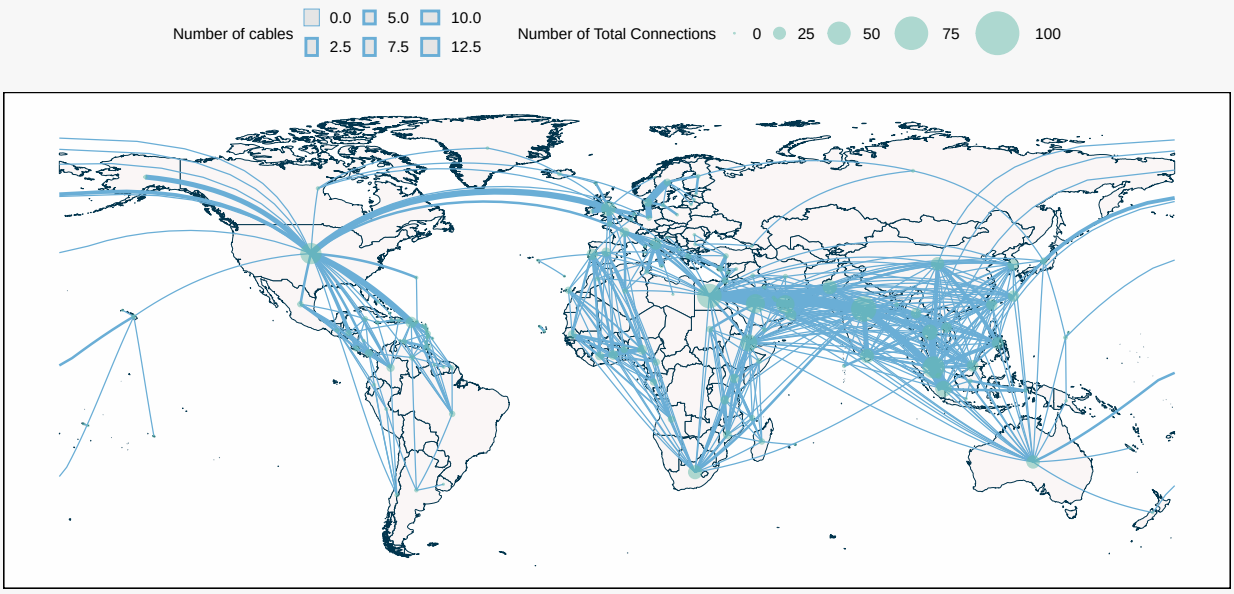
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2011

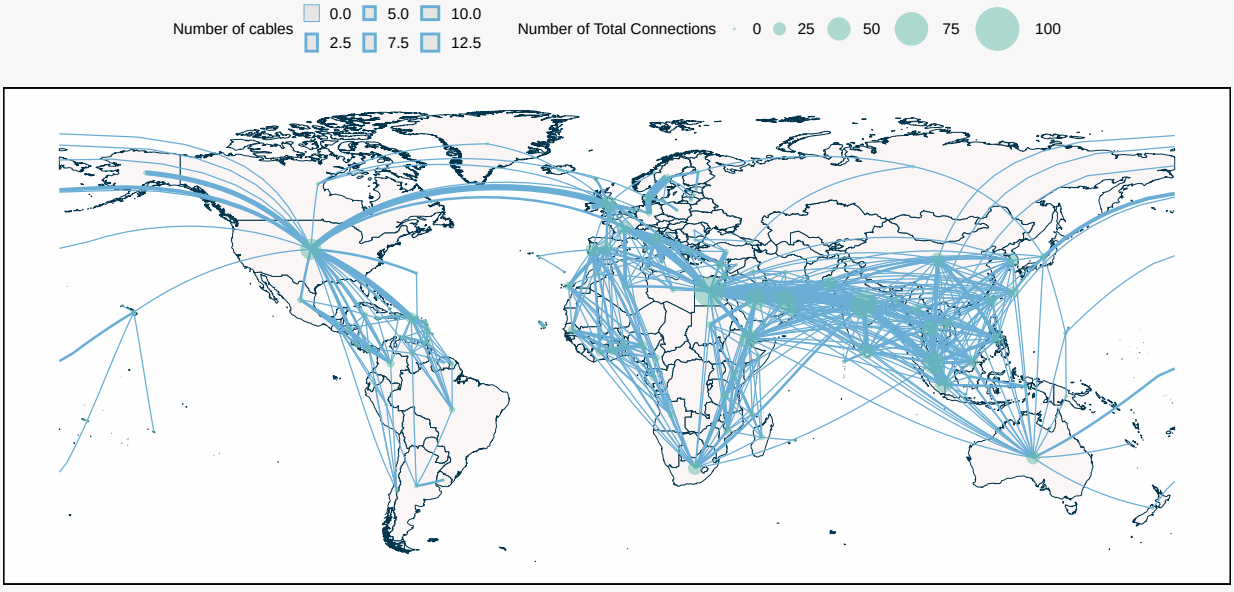
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2012

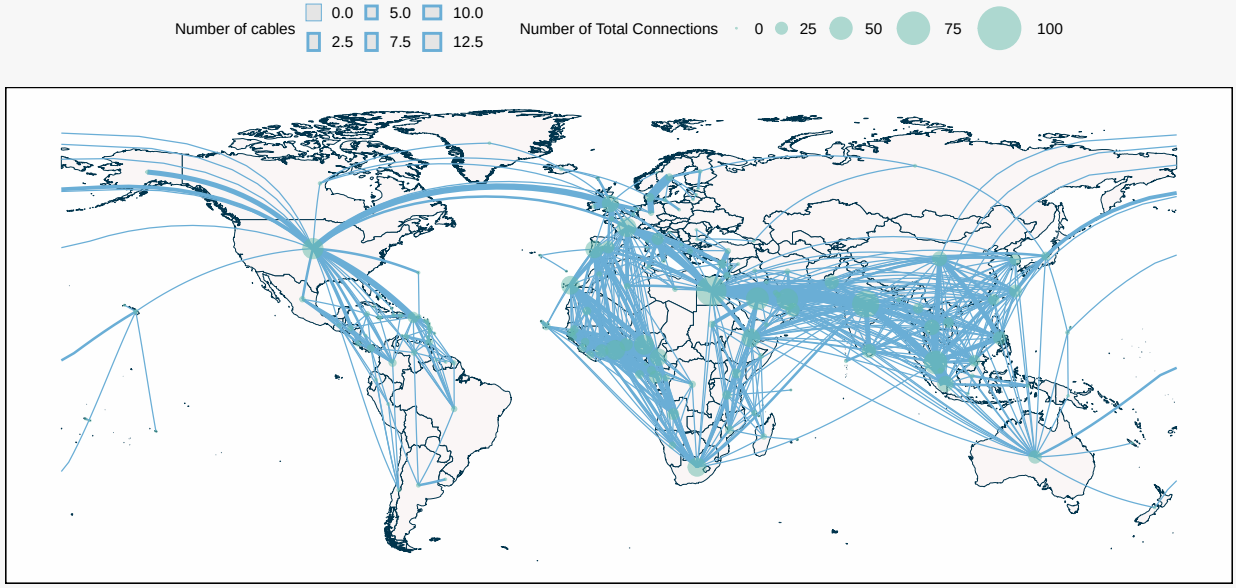
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2013

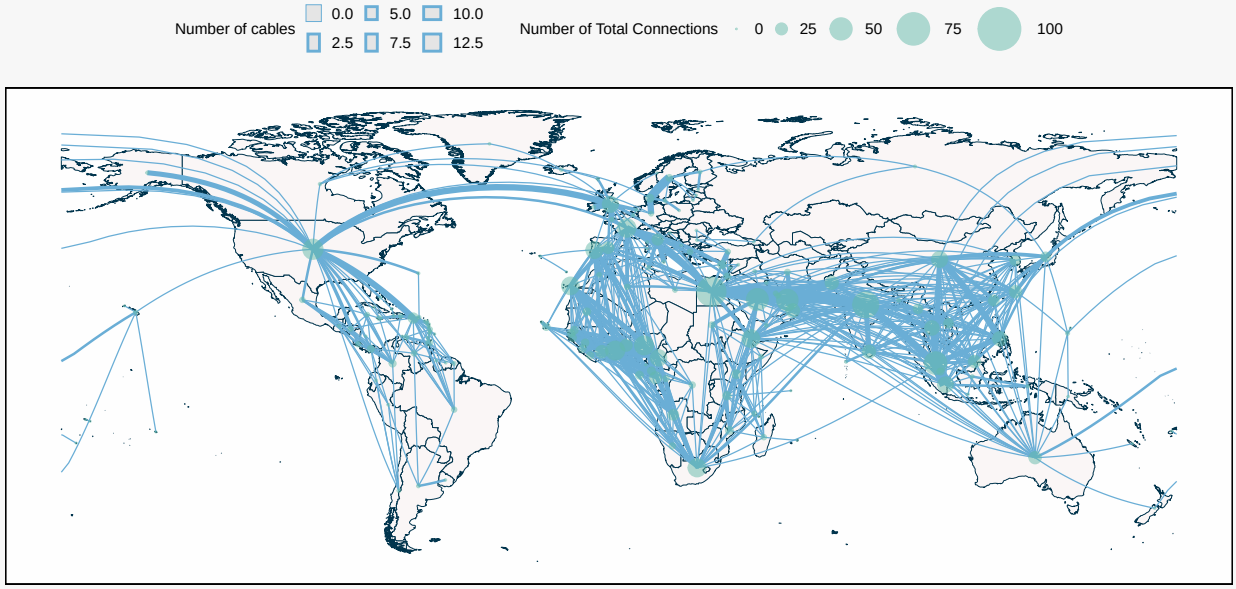
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2014

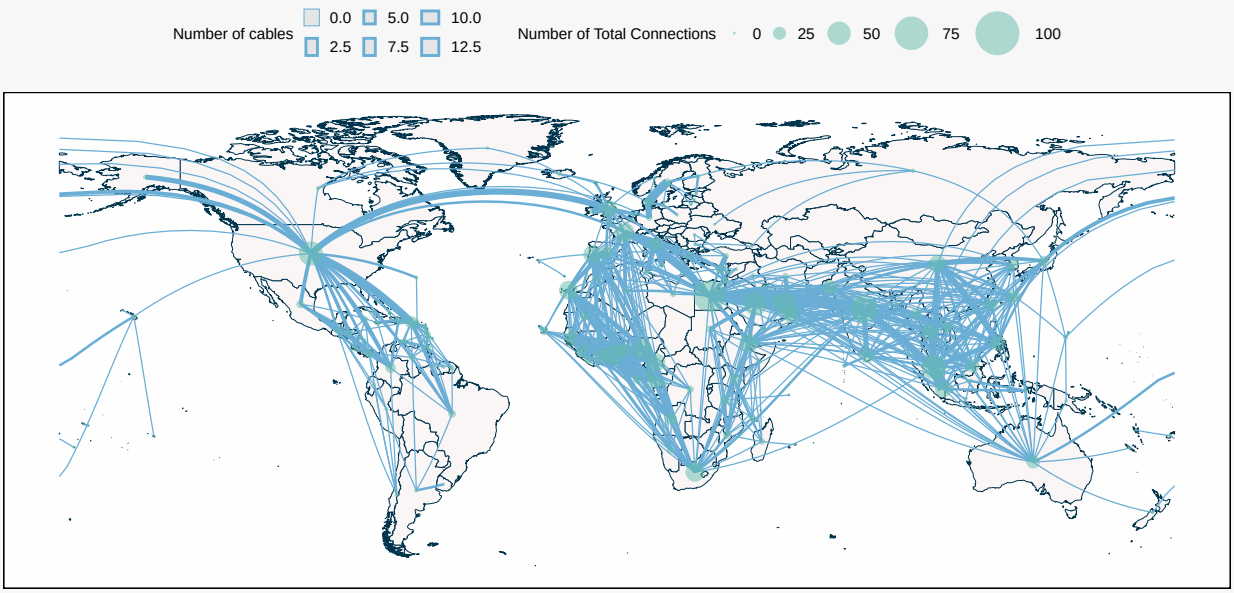
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2015

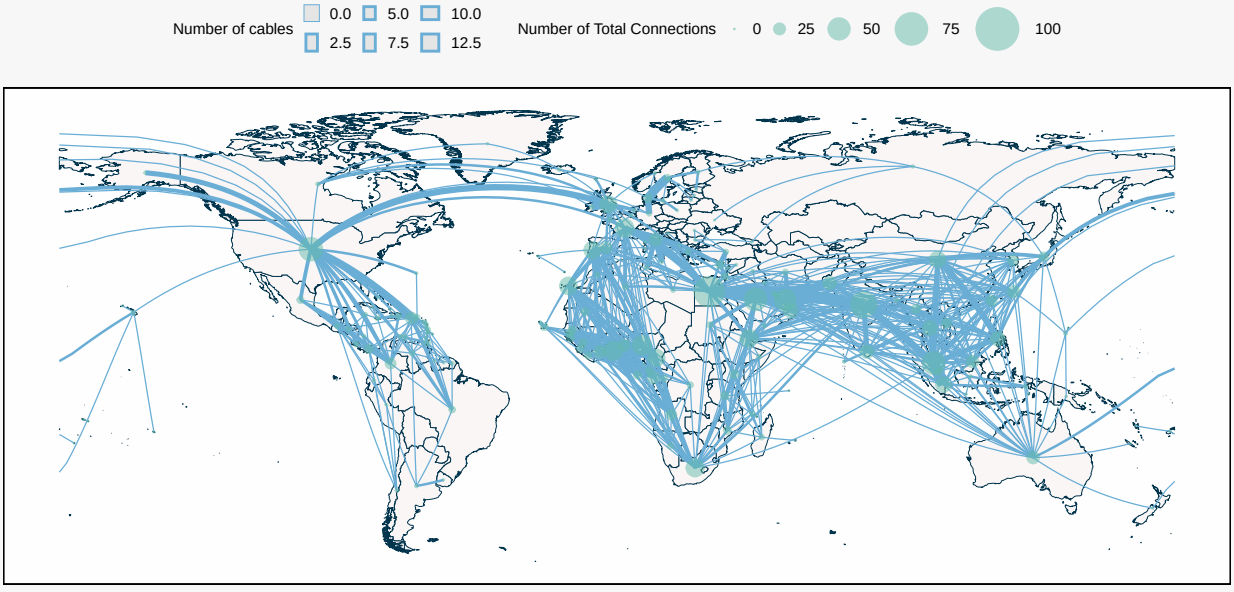
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2016

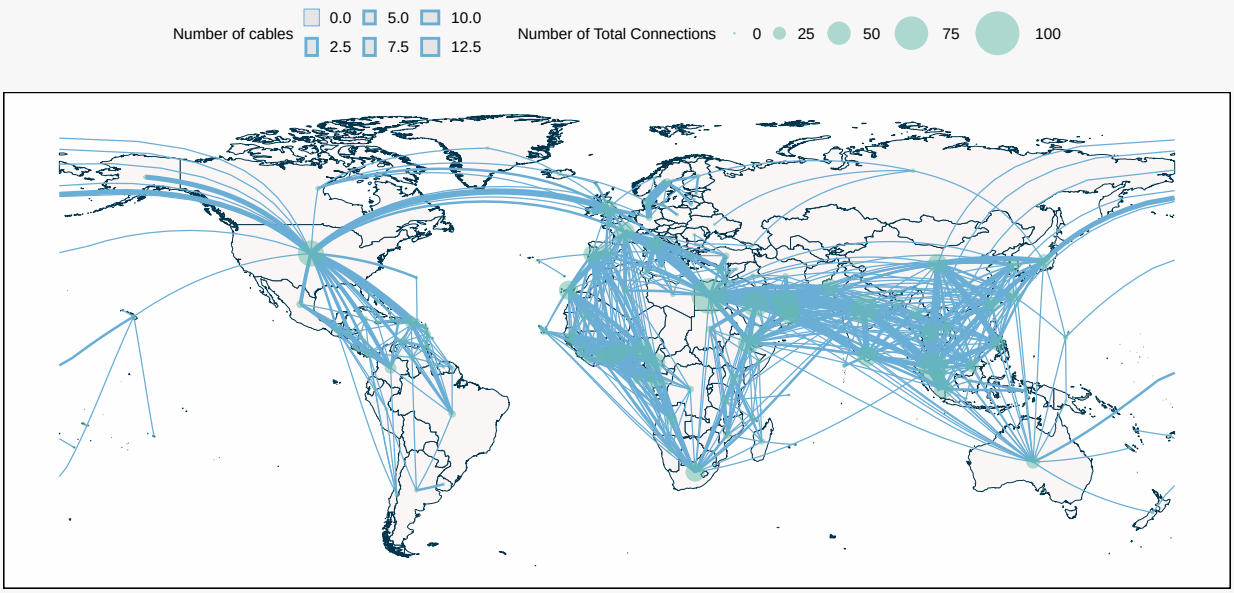
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2017

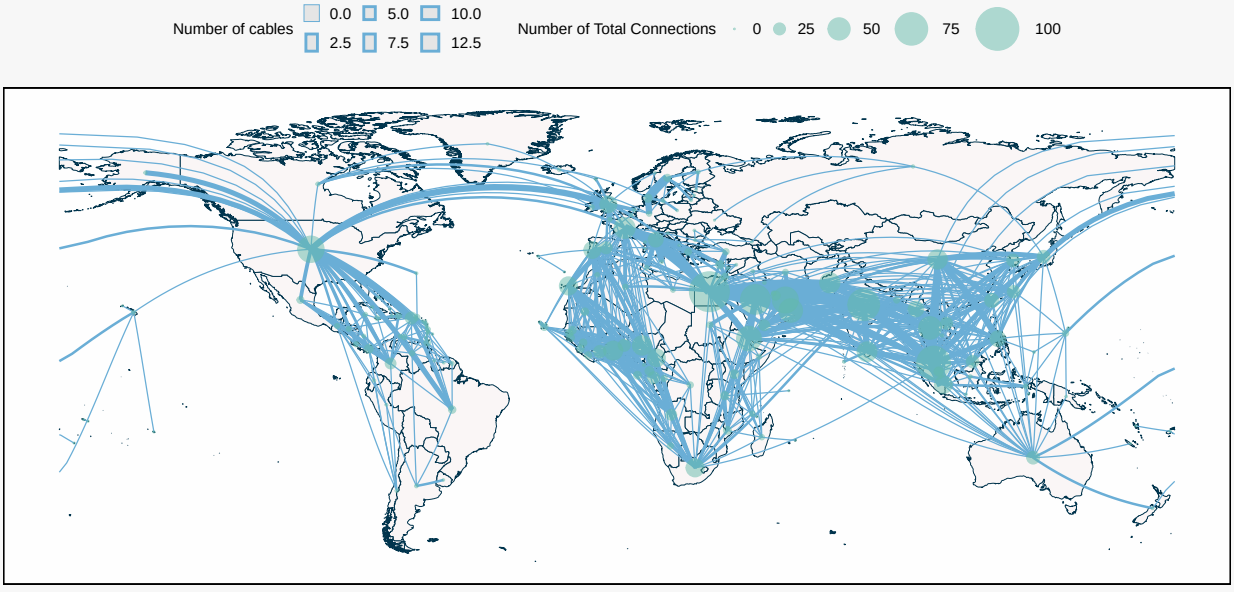
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2018

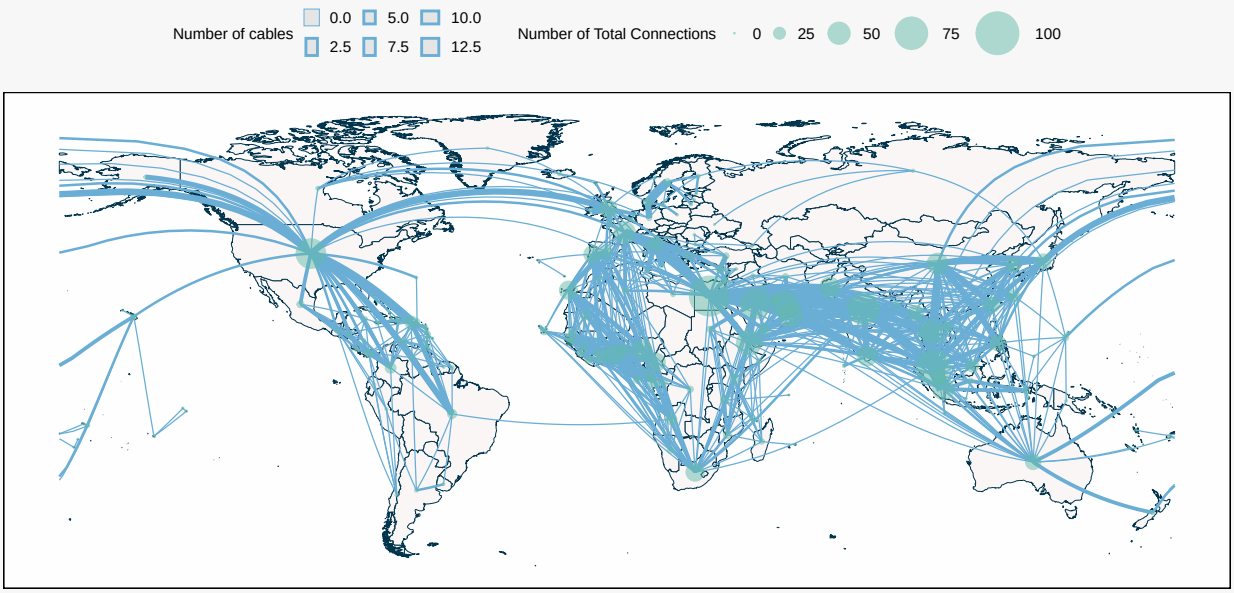
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2019

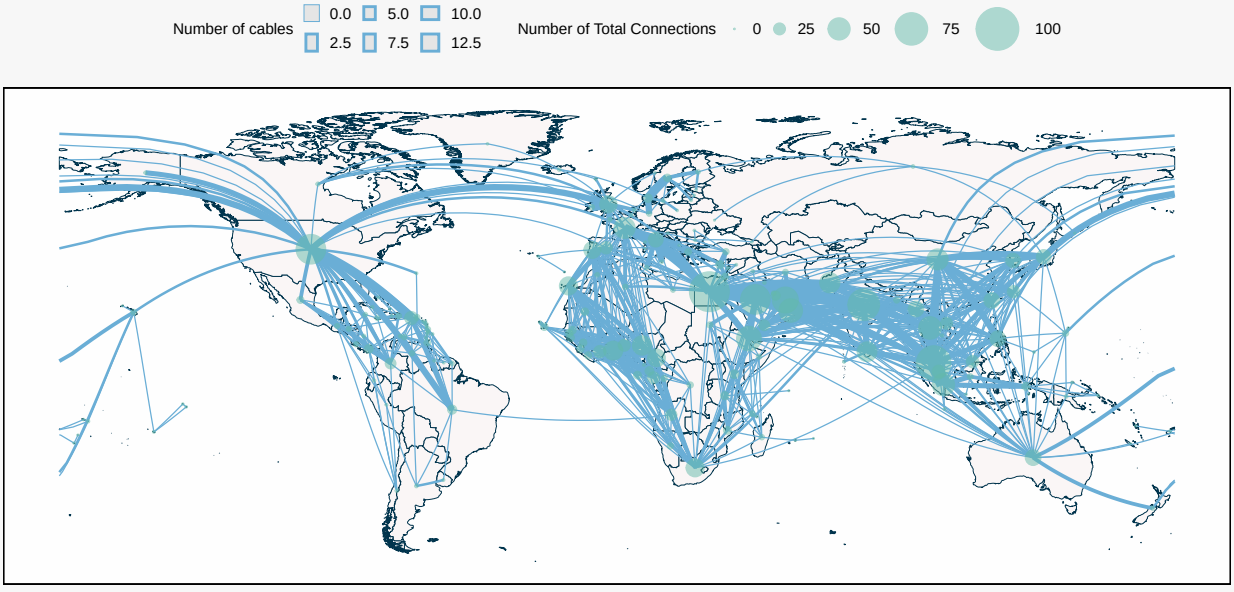
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2020

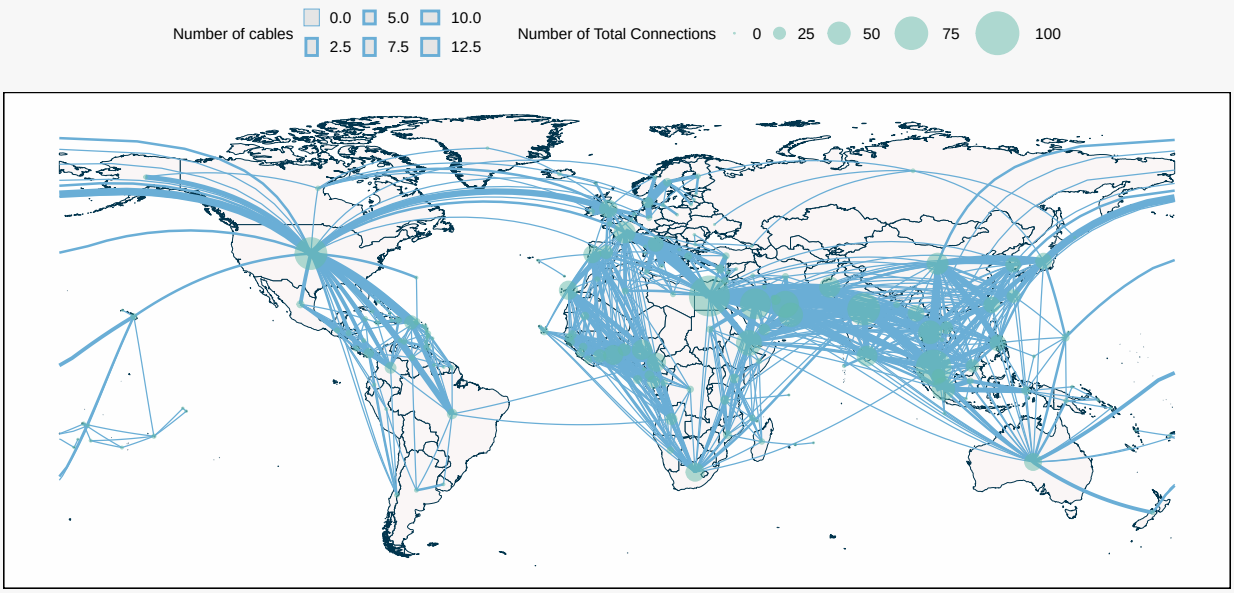
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2021

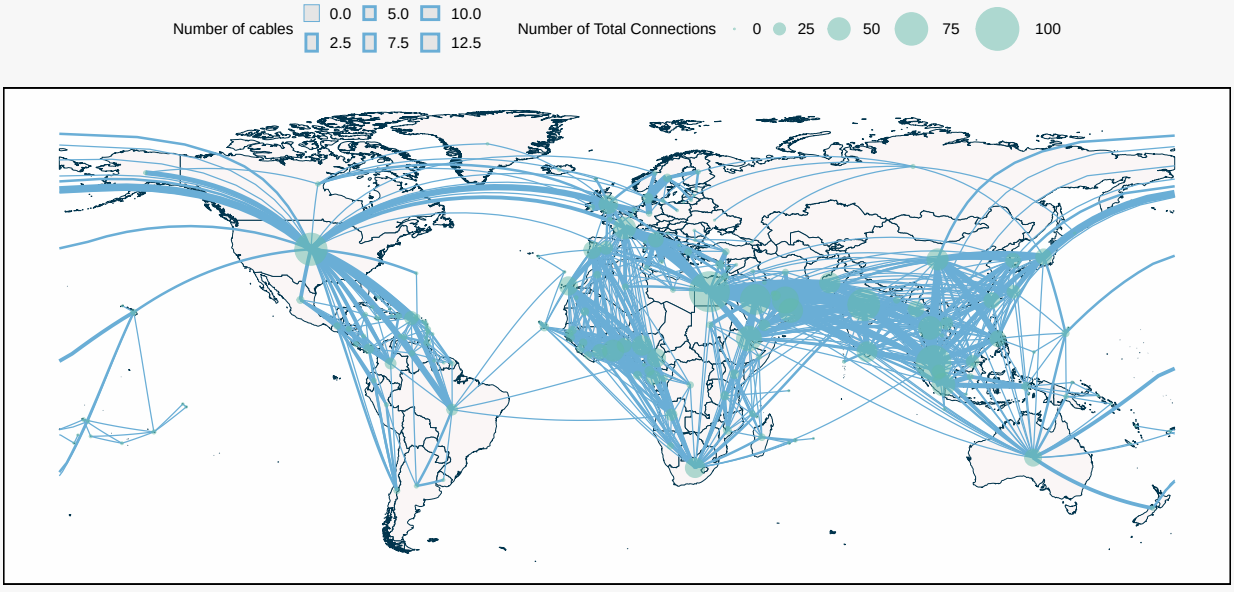
Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography

The World Data Network in 2022

Network Visualization of Subsea Cables and Cloud Datacenters of the main Internet players



Sources: Submarine Cable Map and Cloud Infrastructure Map by Telegeography