iGDB: Connecting the Physical and Logical Layers of the Internet

Scott Anderson
standerson4@wisc.edu
University of Wisconsin
Madison, WI
USA

Loqman Salamatian
ls3748@columbia.edu
Columbia University
New York, NY
USA

Zachary S. Bischof
bischof@gatech.edu
Georgia Institute of Technology
Atlanta, GA
USA

Alberto Dainotti
dainotti@gatech.edu
Georgia Institute of Technology
Atlanta, GA
USA

Paul Barford
pb@cs.wisc.edu
University of Wisconsin
Madison, WI
USA

ABSTRACT

Maps of physical and logical Internet connectivity that are informed by and consistent with each other can expand scope and improve accuracy in analysis of performance, robustness and security. In this paper, we describe a methodology for linking physical and logical Internet maps that aims toward a consistent, cross-layer representation. Our approach is constructive and uses *geographic location* as the key feature for linking physical and logical layers. We begin by building a representation of physical connectivity using online sources to identify locations that house transport hardware (*i.e.,* PoPs, colocation centers, IXPs, etc.), and approximate locations of links between these based on shortest-path rights-of-way. We then utilize standard data sources for generating maps of IP-level and AS-level logical connectivity, and graft these onto physical maps using geographic anchors. We implement our methodology in an open-source framework called the Internet Geographic Database (iGDB), which includes tools for updating measurement data and assuring internal consistency. iGDB is built to be used with ArcGIS, a geographic information system that provides broad capability for spatial analysis and visualization. We describe the details of the iGDB implementation and demonstrate how it can be used in a variety of settings.

CCS CONCEPTS

• **Networks → Network components**.

KEYWORDS

Physical Internet Infrastructure, Logical Internet Infrastructure, Geographic Information Systems

ACM Reference Format:
Scott Anderson, Loqman Salamatian, Zachary S. Bischof, Alberto Dainotti, and Paul Barford. 2022. iGDB: Connecting the Physical and Logical Layers of the Internet. In *ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3517745.3561443

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10…$15.00
https://doi.org/10.1145/3517745.3561443

1 INTRODUCTION

Maps of Internet connectivity can play a vital role in both research and operations. For example, maps of *physical connectivity* (representing PoPs, colocation centers, IXPs, etc. that are connected by fiber conduits[1]) have been used in studies of Internet performance and robustness (*e.g.,* [26, 29, 72]). Physical maps can also be used in operational settings to assess configurations and deployment of new infrastructure (*e.g.,* [29, 30]). Similarly, maps of *logical connectivity* (representing the hop-by-hop IP or AS paths that packets traverse between end points) have been used to investigate a wide range of topics from protocol implementations to traffic engineering to security [16, 33, 39].

The problem of generating maps of physical and logical connectivity has been a focus of research efforts for over two decades. The on-going challenges in this work include the vast size, distributed ownership, dynamic nature, opaque aspects of network device configurations and the sensitivity of service providers. Maps and repositories of physical connectivity have been assembled using various online information sources [27, 67], and the standard approach for generating maps of logical connectivity is through the use of active probe-based measurements [58, 74]. While these maps have been used in many studies over the years, to the best of our knowledge there are no physical and logical maps of Internet connectivity that are readily available, demonstrably up-to-date and consistent across layers.

The objective of our work is to create an open repository of maps of physical and logical Internet connectivity. A key organizing principle of this repository is *cross-layer consistency*, which assures that nodes and links at each layer are related in a way that is consistent with standard Internet organization. That is, nodes in the physical layer house routers operated by service providers that determine both IP and AS level logical connectivity as illustrated in the upper portion of Figure 1. The figure shows an example of the physical-through-logical connectivity associated with a single path between two end points. Physical infrastructure determines the set of possible hops through any individual network, while routing protocols and business relationships determine the specific set of

[1]While our focus in this paper is on wireline connectivity, all of the concepts described can easily be extended to include wireless infrastructure.

hops within and between networks. The lower portion of Figure 1 shows a hand-drawn illustration of the kind maps that we would like to be able to draw and analyze with our envisioned repository.

To create such a repository, several technical challenges must be addressed. These include *(i)* assembling a large corpus of data on locations of physical nodes and links, *(ii)* creating a technique to link physical and logical connectivity, *(iii)* creating a framework for housing, analyzing and visualizing the data and *(iv)* keeping the repository updated on an on-going basis. While these specific challenges are not unique to our study, to the best of our knowledge no prior work has addressed the full set.

We describe a methodology for linking physical and logical Internet maps that aims toward a consistent, cross-layer representation. We posit that *geographic location* is the key mechanism for achieving cross-layer consistency. We build a representation of physical connectivity using on-line sources to identify locations that house transport hardware (*i.e.,* PoPs, colocation centers, IXPs, etc.). We approximate links between these locations based on shortest-path rights-of-way. We then utilize standard data sources for generating maps of IP-level and AS-level logical connectivity, and graft these onto the physical map using geographic anchors.

We implement our methodology in an open-source framework called the Internet Geographic Database (iGDB)[2], which is a system designed to automate the process of collecting Internet topology and measurement data from public sources (as described in Section 2), organize the collected data into a database (Section 3), and enable visualization and analysis through integration with a GIS (*e.g.,* ArcGIS Geographic Information System [6]) as described in Section 4. The repository currently includes 29,220 physical nodes, 8,834 (terrestrial+submarine) links, 7,342 city locations, 210 countries with physical nodes, 102,216 ASNs, 420,913 links between ASNs. For practical reasons, we do not include traceroute measurements in iGDB, but we have a process for including them in analyses and visualizations. We combine information from the best available public sources, but do not claim that iGDB is a complete database on all logical and physical Internet topology data. Rather, we present iGDB as an open repository for the Internet research and operations communities, with the goal of accepting and integrating community contributions to ensure its continued accuracy and relevance.

We demonstrate the utility of iGDB through a series of use case examples. At the highest level, there are three basic use cases for iGDB, which include *(i)* a unified repository for Internet connectivity data that would otherwise be difficult to assemble from disparate sources, *(ii)* geo-spatial analysis of Internet connectivity including other data types available in GIS shapefile format and *(iii)* generating (potentially complex) visualizations of Internet connectivity. The specific use case examples that we provide in Section 4 include assessing the geographic footprint of autonomous systems, identifying physical paths associated with logical path (traceroute) and network measurements, and inferring location information from logical measurements. In each case, we provide examples of visualizations that would be challenging to produce without iGDB.

[2] The iGDB code may be found at: https://github.com/standerson4/iGDB.

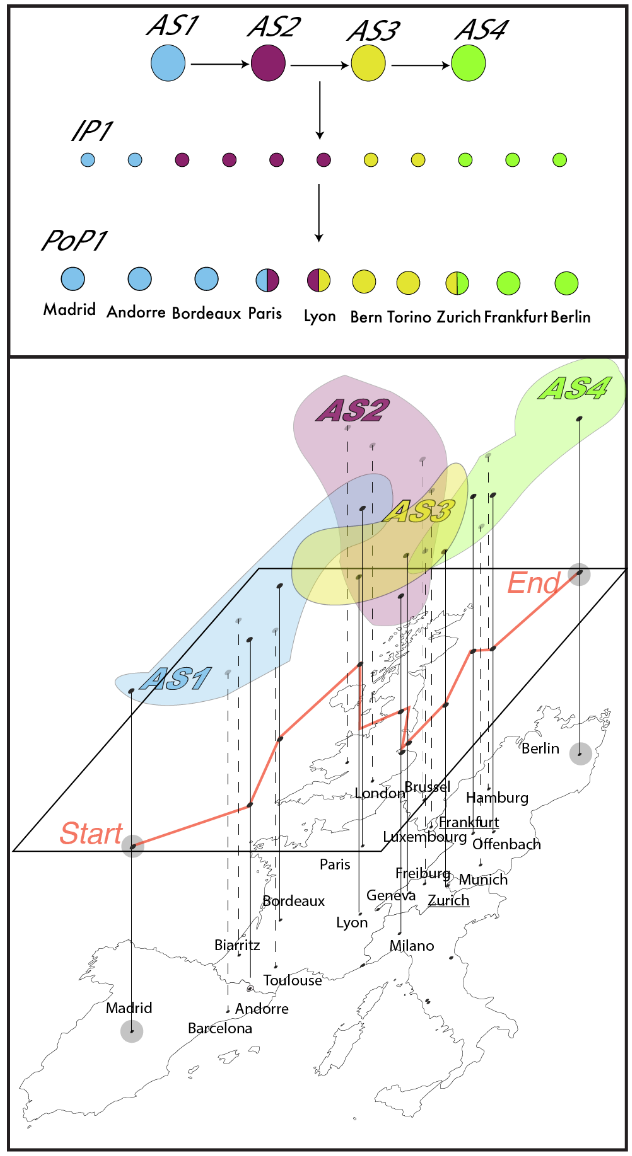


Figure 1: Illustration of a consistent representation across physical and logical layers of a network path between Madrid and Berlin.

2 INPUT DATASETS

iGDB is motivated by the observation that there are currently many high-quality datasets published by both researchers and network operators on Internet topology. Due to the fact that the creators or collectors of these datasets are often motivated to address a unique set of problems related to Internet topology, each dataset contains different subsets of information formatted for a specific audience and purpose. Some datasets focus on documenting physical infrastructure, such as Points of Presence (PoPs), Internet Exchange Points (IXPs), and fiber optic cables [27, 44, 76]; while other datasets focus on creating a corpus of logical infrastructure, such as

IP addresses, Autonomous Systems (ASes), and AS interconnection relationships [40, 42, 61].

The objective of iGDB is to provide researchers and network operators with these datasets in an accessible toolkit that allows them to gain new insights on Internet topology by bridging the gap between logical and physical topology information. Collecting information from different sources in different formats requires care and a deliberate methodology. iGDB is designed to automatically collect snapshots from each of the sources described below. In order to capture changes in each dataset over time, iGDB saves timestamped snapshots of each source, then automatically processes and loads the data. Additionally, as researchers make new datasets available to the research community in the future, we anticipate providing options to integrate them, either ourselves or through community contributions.

In the remainder of this section, we describe the information we gather from each of these datasets to provide a more comprehensive and joint view of physical and logical Internet topologies. We then describe how we organize the collected information in the iGDB architecture in Section 3.

Internet Atlas (Physical Infrastructure): Internet Atlas was an academic study with the goal of consolidating geographic information about Internet PoPs and fiber optic cable infrastructure to enable continuing research on the Internet's physical topology [27]. Internet Atlas includes maps of physical connectivity from over 1.5K networks worldwide. We used this dataset to create a repository of PoPs that are categorized by the owning organization and physical location. Because many service providers consider revealing the exact physical paths fiber optic cable traverses between nodes a security risk, we describe a technique to infer approximations of the physical paths connecting nodes in Section 3.1. For a review of the accuracy, precision, and completeness of Internet Atlas, we point the reader to the original research, in which the authors described the steps they took to *verify* the Internet Atlas data, *i.e.,* to ensure that the entered data is an accurate representation of the original data source; how they *validated* the data, *i.e.,* demonstrated that the data accurately reflects the real world; and how they determined the *completeness* of their data, *i.e.,* how much of all Internet infrastructure is included in Internet Atlas [27].

Telegeography (Physical Infrastructure): Submarine cables are conduits for international data transfer and play a key role in providing digital services for end users by connecting many distributed users to the centralized data centers from which Internet services are provided [18, 47]. Although critical for connecting transoceanic populations, submarine cable infrastructure was not included in the Internet Atlas study, so we collected data from an alternate, openly available source, Telegeography [76]. The data we imported includes the consortium of companies overseeing each cable, the cable segment physical paths, and their associated landing points.

PeeringDB (Physical and Logical Infrastructure): PeeringDB is a site where service providers publicly announce their presence at public and private community peering locations with the goal of establishing BGP peering relationships with others to connect users with services [61]. At the physical layer, PeeringDB provides information on the physical locations (*i.e.,* addresses, as well as lat/long coordinates) of the nodes where interconnection takes place. On top of the PoP-level information, PeeringDB also includes information for more than $24K$ networks[3] that share information regarding their physical footprints and peering policies. We also gather IP prefixes corresponding to IXPs by extracting information from the public facilities fields. In their research, Lodhi *et. al* describe not only what data is included in PeeringDB, but also how representative that data is across business types and geography, as well as how complete and current the data is [48]. Additionally, in private conversations we had with PDB developers and network operators, they emphasized that PDB conducts internal audits to identify inconsistencies and verify the correctness of the database.

Packet Clearing House (Physical and Logical Infrastructure): Packet Clearing House (PCH) is a non-profit created to facilitate the building and support of Internet Exchange Points and maintains a directory of all Internet exchanges worldwide in an IXP directory [40]. We augment the peering facilities and networks identified in PeeringDB with information on IXP names and ASes from PCH.

Hurricane Electric (Physical and Logical Infrastructure): Hurricane Electric (HE) is a worldwide transit ISP with connections to more than 7K networks across 200 exchange points [42]. We adjoin the HE Internet Exchange Report to the dataset of IXPs and the set of IXP IP prefixes.

EuroIX (Logical Infrastructure): EuroIX collects data directly from IXPs through a recurring automated process, providing a comprehensive public source of IXP-related data directly originating from the IXPs [32]. In contrast, PCH and PeeringDB's data is compiled manually by the ASes and the IXPs themselves, which may be more error-prone and at risk of being of out of date.

Reverse DNS (rDNS) look-up (Logical Infrastructure): We collect the publicly available IPv4 PTR lookups for IPv4 addresses from Rapid7 [63].

AS Rank (Logical Infrastructure): We use the public BGP routing data returned by the CAIDA AS Rank API. The resulting AS topology returned is the aggregation of all the RouteViews [68] and RIPE RIS [65] BGP announcements for the first 5 days of a month. It comprises a graph with undirected edges between two ASes if two ASes were adjacent in an observed AS Path.

RIPE Atlas Anchor Meshes (Logical Infrastructure): RIPE Internet Atlas is an Internet measurement platform with small probes installed in networks around the world that take periodic measurements, such as traceroute and ping, to remote servers and other RIPE Internet Atlas probes [66]. In addition to these measurements, each probe has an associated IP address, ASN of the network that hosts the probe, as well as the approximate geographic location of the probe. We include this information as it links logical data (ASN) with physical data (geographic location of each probe), creating an important connection between the two layers. Additionally, we use traceroute measurements when identifying possible physical paths of logical measurements, which we discuss in Section 4.2. To understand the inconsistencies and biases from the RIPE Atlas network, Bajpai *et. al* describe the geographic and AS distribution of RIPE Atlas nodes, the bias toward deployment by technically-inclined volunteers, and a comparison of measurements by probe hardware type [17].

[3] Collected in February 2022.

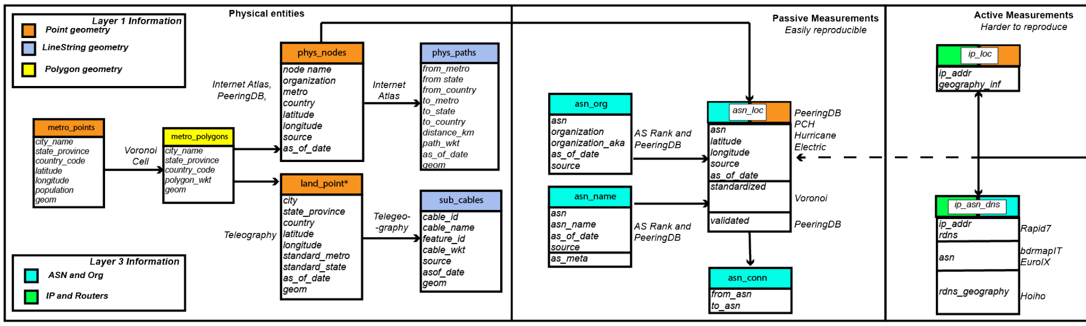


Figure 2: Schematic view of iGDB.

Summary: Each data source described above provides important insights on Internet architecture and topology. iGDB provides a baseline database that uses geographic information to make connections between these various sources. As these data sources are updated over time, users can refresh their local data as frequently as required. iGDB is able to readily incorporate changes by obtaining new snapshots and automatically processing and importing the updated data.

3 DATABASE ARCHITECTURE

Each dataset described in Section 2 provides important information on Internet topology. We posit that consolidating these datasets can provide new insights for network researchers. However, doing so is challenging because the datasets are published by different sources in a variety of formats with diverse nomenclatures. One of the contributions of iGDB is to unify the conventions of these datasets in a logical fashion and make them available to the research and operational communities.

We chose to use a relational database model to organize the data because it integrates easily with ArcGIS and is well-established for storing and querying information. Throughout the process of designing the database structure, we made deliberate decisions to incorporate appropriate attributes to link the physical and logical views of the Internet.

As introduced in Section 2, we divide the collected information into two broad categories: relations about physical attributes and relations about logical attributes. We chose this approach for two reasons: *(1)* network researchers are familiar with the concept of a layered architecture and *(2)* data from each source mostly falls into one of these categories. Geographic information is stored in relations about physical structure, but, because geographic location is our unifying characteristic, it is important for us to be able to query across all relations to provide a geographic perspective on the logical structure of the Internet.

The structure of the Internet changes as new physical infrastructure is built, old infrastructure is replaced, peering relationships are established or removed, device configurations are modified, *etc.* Physical topology changes are relatively slow due to the time required to establish and integrate new physical infrastructure. Logical topology changes may occur much more frequently, sometimes through automated processes such as when routing protocols adapt data flows to topology changes. Some researchers may be interested in a single snapshot of the topology, for example the most recent view available, but others may require a better understanding of topology and how it changes over time. To enable both types of actions, iGDB includes an *as of* date as an attribute for all collected data. The *as of* date may be included in user queries as necessary.

The relational structure of iGDB is shown in Figure 2. We describe the architecture of the physical and logical elements of iGDB in Sections 3.1 and 3.2. We then describe how we bridge the logical and physical elements in Section 3.3.

3.1 Layer 1: Physical

Our model for the physical structure of the Internet is a network where edges are the physical cables that connect computer systems; nodes are physical interconnection facilities (PoPs, IXPs, and cable landing sites where cables from different organizations are physically connected with network devices (routers and switches) for logical information transfer. We collected data on physical nodes from Internet Atlas, PeeringDB, and Telegeography. We collected data on physical edges from Internet Atlas and Telegeography.

We addressed three main challenges with this data: *(1)* completeness - because there are so many Internet service providers across the world, it is difficult to collect information from all of them; therefore we used multiple sources of information available: Internet Atlas, PeeringDB, and Telegeography; *(2)* the exact paths that physical cables traverse is unknown to outside researchers, so we appeal to research from Durairajan *et. al* who observed that long-haul cables follow rights-of-way along existing networks such as roadways, rail, and power lines [29]; *(3)* the locations of nodes may be ambiguous because of non-standard naming conventions.

We overcame inconsistencies in location through the location standardization process described next. To standardize location names, we use the following process to spatially map each node to the closest urban area from a single data source of urban areas. We use as a single source of urban areas a shapefile of global populated places from Natural Earth [5]. This point shapefile includes 7,342 urban areas of the world, including major cities, populated towns, villages, and other areas with regional significance. To map each node to the closest city, we use ArcGIS to divide the entire Earth into a set of 7,342 Thiessen polygons [7] that enclose the urban areas from the Natural Earth shapefile, as shown in Figure 3. Any point inside each of these Thiessen polygons is geographically closest to the single urban area used to create the polygon.

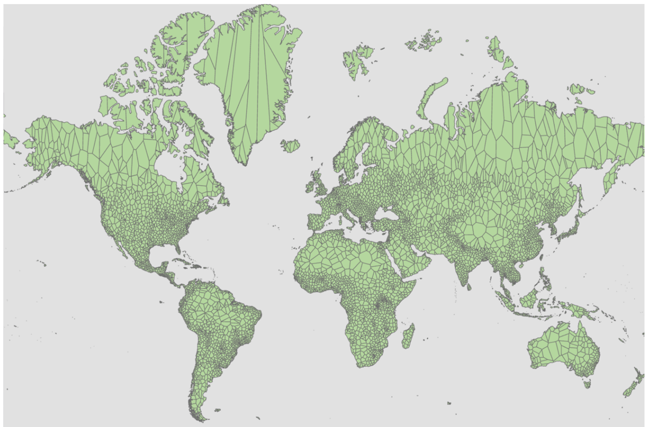


Figure 3: Map of the world divided into 7k Thiessen polygons around urban areas with mask over oceans.

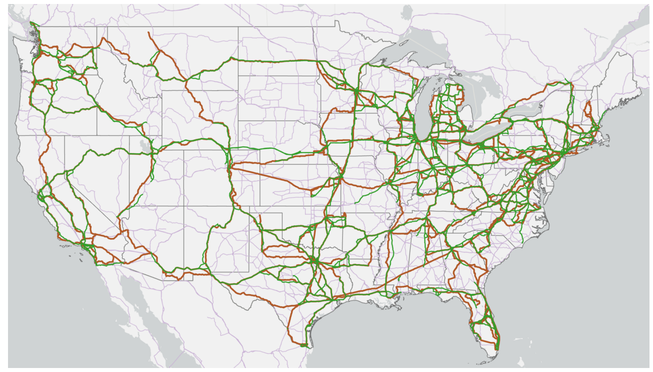


Figure 4: Recreation of the InterTubes long-haul fiber optic cable routes (in brown), the iGDB shortest-path routes that most closely (within 25 miles) follow the InterTubes long-haul links (in green), and the iGDB shortest-path links that do not closely follow the InterTubes links (in purple).

After defining this information, we use the relation *city points* to store the name and location of the standard cities we used to build the Voronoi diagram of the Earth's surface. The *city polygons* table contains the physical geometry for the area around each of the city points, stored as a polygon in the well-known text (WKT) format [77]; the polygons relate to the city points by city, state, and country name.

With the Earth divided into polygons surrounding the urban areas, we next map the network nodes to a standard location name. Each node has a latitude and longitude location associated with it from the original data source. We use ArcGIS to spatially join the network nodes with the Thiessen polygons to associate each node with the closest urban area. After this operation, each node has a new attribute with a standard city name and location. This allows us to easily compare all network nodes by location, even when each node is originally labeled with a non-standard naming convention.

With physical nodes thus standardized, we organized the data into eight relations, as seen in Figure 2. There are five relations dedicated to physical nodes: *ASN location*, *city points*, *city polygons*, *landing points*, and *physical nodes*. There are two relations describing physical edges: *standard paths* and *submarine cables*.

We divide physical nodes into multiple relations to group them by the type of physical node. The *ASN locations* relation maps ASNs to locations, which are derived from the ASNs listed as having a peering presence at an IXP or a private peering facility from PeeringDB. The *physical nodes* relation includes the node names, owning organization, and location for PoPs from Internet Atlas as well as the PoPs and IXPs in PeeringDB.

The *submarine landing points* table includes the physical locations, including name, lat/lon, and standard city name, where submarine cables are brought on shore. There may be multiple submarine cables that land at each landing point. The *submarine landing points* table relates each of the submarine cables to the specific landing points where it comes on shore. This data is collected from Telegeography.

Physical edges are divided into two relations because of the different attributes that describe terrestrial long-haul cables and submarine cables. Both relations store the physical edge path as a LineString or MultiLineString WKT that can easily be converted into a geometry by a GIS. The data in *submarine cable paths* is collected from Telegeography and that relation stores the cable name and physical cable path. *Standard paths* are derived from known rights-of-way between nodes that have a physical edge described in Internet Atlas in the following manner.

We imported the PoP-level connections from Internet Atlas. However, we do not use the exact cable paths from Internet Atlas, which are not made available due to their sensitive nature as illustrated by the fact that fiber conducts are primary targets in case of disruption efforts [75] and the decrease in the number of ISPs providing exact maps of their connectivity. We only import the fact that two physical nodes are physically connected. Without knowledge of the exact physical path the fiber optic cable takes between the nodes, we generate an approximation of the physical path. As observed in earlier work, long-haul cables often follow rights-of-way along major roadways [27, 29]. In addition to academic research that demonstrated fiber optic cable following existing rights-of-way, many countries and industry groups have laws and standards codifying how telecommunications infrastructure should be placed alongside other critical infrastructure with defined rights-of-way [2, 3, 14]. Finally, a visual inspection of published ISP infrastructure maps, such as those found in [8, 11, 12] demonstrate how important these road and rail right-of-way laws are to enable an ISP to cost-effectively deploy global telecommunications infrastructure. Additional right-of-way networks (rail networks, natural gas pipelines, etc.) may be considered through community contributions to iGDB to extend path diversity. We use this information on existing road networks to generate an approximation of the physical path the fiber optic cable connecting the two nodes follows. This is accomplished by determining the shortest route connecting city pairs along the right-of-way network, which can be generated with a geographic information system (GIS). We store the resulting physical path using the LineString WKT between the two nodes and classify each physical path with the source and destination standard city names [77].

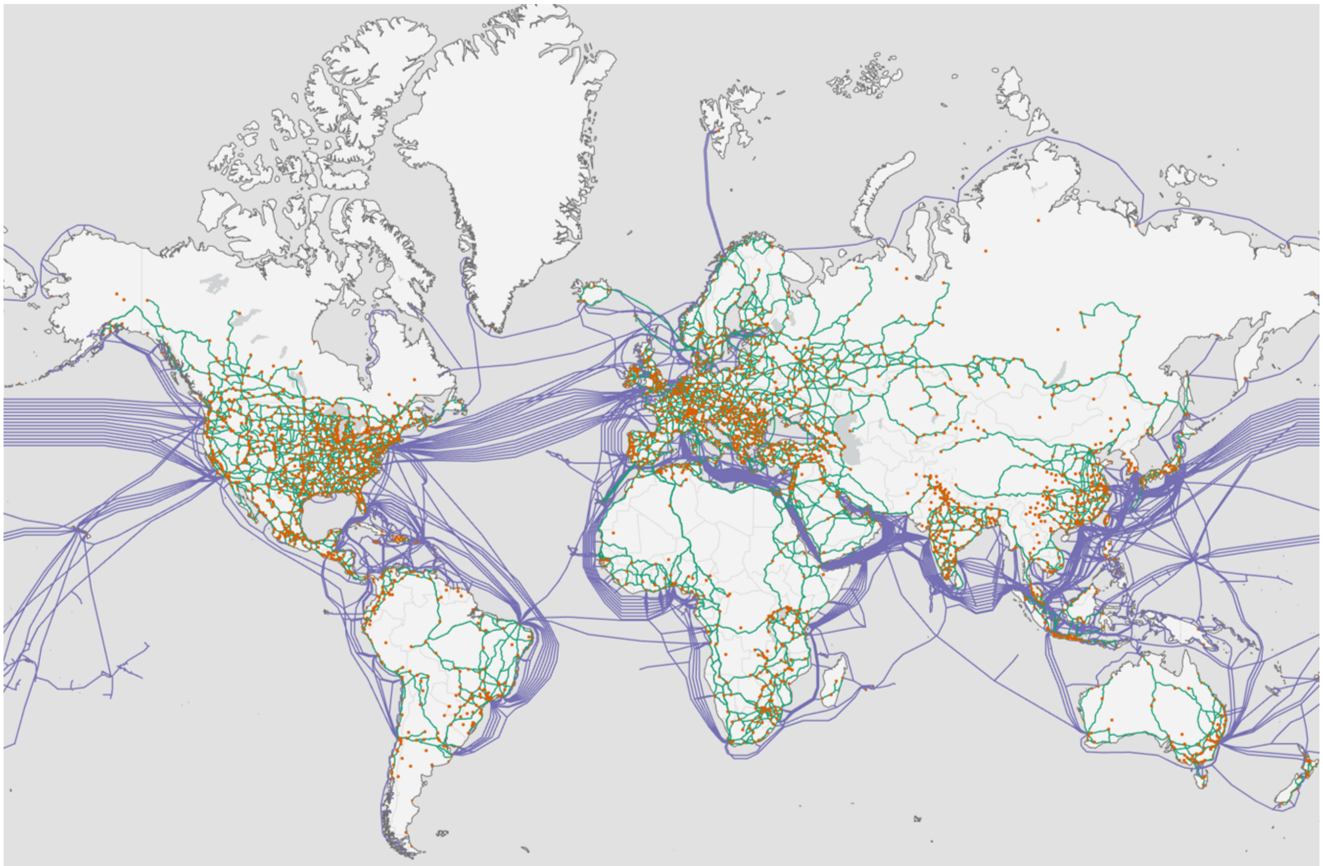


Figure 5: The physical elements of iGDB include nodes from Internet Atlas, PeeringDB, and Packet Clearing House (PoPs, IXPs, etc.) in orange, physical paths inferred to follow established rights-of-way alongside vehicle transportation infrastructure in green (unconnected nodes result from incomplete information on physical connectivity), and submarine cable physical paths from Telegeography in purple.

To demonstrate the benefit of using shortest-path routes along transportation networks to approximate the actual paths that physical fiber optic cables trace, we conducted a qualitative spatial comparison of the iGDB fiber optic network with a recreation of the InterTubes long-haul map. A visualization of this comparison is shown in Figure 4. From this spatial comparison we make three observations. First, most of the InterTubes fiber optic cables are closely approximated by the iGDB shortest-path links. However, we do not claim that all links are approximated. For example, the InterTubes link in the southeast US – in brown – from Altanta, GA to Houston, TX that is not approximated by any iGDB link most likely follows a natural gas pipeline that was not included as a potential right-of-way pathway when developing the iGDB shortest path routes. We discuss how to include additional right-of-way information into the iGDB shortest-path routes in Section 5. Second, the iGDB links do not exactly follow any of the fiber optic cables from InterTubes. This is by design, as the iGDB shortest-path links are intended to closely approximate the real-world fiber optic cable paths, without mirroring them exactly, for security purposes. Finally, iGDB includes (in purple) many potential alternate paths along transportation networks that did not have long-haul links at the time the InterTubes map was developed. We cannot confirm that these transportation network paths do not currently have fiber optic cable infrastructure alongside them, but their existence means that there may be transportation corridors for ISPs to expand their networks to provide service to more people and/or improve the redundancy of their networks. With inside knowledge of an ISP's network, iGDB provides the information for an ISP to expand or improve their network.

Figure 5 is a visual representation of the physical nodes and edges included in iGDB. While this map is similar to maps of physical Internet infrastructure from Internet Atlas [27], Infrapedia [9], or the ITU transmission network map [10], we emphasize two important distinctions: *(1)* the map generated by iGDB and the data used to generate the map is openly available and can easily be recreated or extended by adding data from the other sources and *(2)* this map displays information at the physical layer in iGDB, which can be fused with information on logical layer attributes, as described in Section 3.2.

3.2 Layer 3: Logical

Our model for the logical structure of the Internet is also a network graph, but the nodes and edges are defined differently. In our coarse-grained logical model, nodes are ASNs and edges are the interconnection relationships between ASNs. We define three relations with this logical information: *ASN-AS name*, *ASN-organization*, and *ASN-connection*. The *ASN-AS name* table relates the ASN to the AS name, as collected from ASRank and PeeringDB. Because we gather information from multiple sources, each ASN may match multiple AS names. The *ASN-organization* table similarly maps ASN to organization names from ASRank, PeeringDB, and PCH. We do not attempt to standardize AS names or organization names from different sources. Finally, *ASN connections* defines the ASN to ASN interconnection relationships as collected by ASRank. We preserve inconsistencies between AS names and ASNs, as well as between organizations and ASNs, from their original sources because the user may encounter different naming conventions during their research. When conducting analysis, inconsistencies may be minimized and accounted for using appropriate SQL queries. Each relation includes the ASN entity, which is a unique identifier that may be used to relate various AS names and organization names.

We can illustrate the importance of using ASN as a key between AS name and organization with AS2686. AS2686 has an AS name of "ATGS-MMD-AS" from ASRank (extracted from WHOIS) and of "as-ignemea" from PeeringDB (extracted from an IRR database). Similarly, AS2686 has an organization name of "AT&T EMEA - AS2686" from PeeringDB, "AT&T Global Network Services Nederland B.V." from PCH, and "AT&T Global Network Services, LLC" from ASRank (WHOIS). By maintaining two relations, *ASN-AS name* and *ASN-organization*, we capture the inconsistent names that AS2686 is known by while maintaining a common key (2686) relating the various entries that result when collecting similar data from multiple sources.

While inter-AS topology connectivity provides a coarse view of the logical connectivity, some use cases (*e.g.,* those described in Section 4.2 and Section 4.4) require geographically mapping the logical topology between IP addresses. Incorporating IP topology into iGDB requires three preparatory steps: *(1)* mapping each IP address to ASN (*e.g.,* using bdrmapIT [51]), *(2)* mapping each IP address to FQDN (*e.g.,* using rDNS or Rapid7 [63]), and *(3)* geolocating each FQDN (*e.g.,* using Hoiho [50]). We completed this process for IP addresses from RIPE Atlas anchor traceroute mesh measurements using bdrmapIT for IP:ASN mapping, Rapid7 for IP:FQDN mapping, and Hoiho for FQDN:geolocation. We store this data in the *IP-AS-DNS* table. However, we anticipate that users will want to modify this table with additional IP address mappings or by using different techniques or datasets for address translation, name resolution, or geolocation. We therefore provide the ability to add new user-generated mappings to the iGDB database.

3.3 iGDB Table Relationships

A key contribution in our work is to devise bridges linking the logical and geographical tables. We first create a table to capture mapping between AS and PoPs called *ASN-Locations* as inferred from PeeringDB, PCH and Hurricane Electric. We translate the raw city to metros via the tessellations obtained from the Thiessen polygons. This table models the geographic footprint of an ASN. We also allow for the option to push the geographic locations and ASN entries inferred from the *IP-ASN-DNS* to the *ASN-Locations* table.

Table 1: Select database characteristics.

Type	Value
Number of ASes	102,216
Number of organizations	81,879
Number of physical nodes	29,220
Number of countries with nodes	210
Number of inferred physical paths	8,323
Number of submarine cables	511

The main challenges that we face with this mapping are: *(1)* IP to AS mapping and *(2)* remote peering. IP to AS mapping is problematic because a link between two ASes is usually assigned IP addresses from one of the ASes. As a result, mapping the IP address to the AS announcing the smallest subprefix can result in wrongly inferred ownership of links. Furthermore, traceroute probes are known to result in a wide-array of behavior and it is impossible to know a-priori what interface responded to the probe increasing complexity in identifying ownership of the hops. To address this challenge, we leverage bdrmapIT [51], a state of the art technique to map *network borders*. In regard to (2), remote peering is a service provided by IXPs that enables the virtual presence of an AS in the IXP for a smaller fee than a full physical presence and may be executed from PoPs across the globe [21]. However, remote peers are not distinguished from those physically present making their physical locations ambiguous. We acknowledge that remote peering exists and include a flag in the *ASN-Locations* to inform when we believe an AS is virtually located in a physical infrastructure. We classify ASes as remote peers using the technique described in [57].

After organizing and standardizing the information collected from various sources as described above, iGDB consists of the relations depicted in Figure 2 and contains information as described in Table 1.

Critical to making iGDB useful to outside researchers and operators was to ensure that the individual relations could be queried in tandem to elucidate meaningful relationships. In the physical-layer tables, we accomplished this with the geographic standardization of node locations and physical edge endpoints. At the logical level, we used ASN as the key to relate the various AS names and organization names from different sources.

There are two attributes available to join the physical and logical views of the Internet topology: ASN and organization name. Physical nodes are all owned by an organization and the 'organization' attribute may be used to relate back to logical ASNs. Similarly, ASN physical peering locations are defined in the *ASN location* table at the physical level and ASN can be related to AS name or organization at the logical level. Because the data is organized in this fashion and we have established standardized keys that relate physical and logical information, we can use iGDB in a variety of research and operational settings, as described in Section 4.

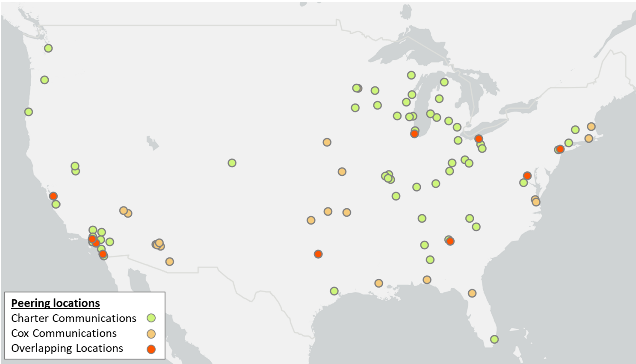


Figure 6: Metro areas with with a peering presence from Charter Communications (green) and Cox Communications (orange) in the US. Metro areas with a presence from both ISPs shown in red.

4 EXAMPLE USE CASES

Using the database generated by iGDB, which fuses physical and logical topological information about the connectivity structure of the Internet, researchers can quickly and more easily extend their research in new directions. Though we anticipate that researchers will use iGDB in many ways that we cannot predict, in this section we provide a series of examples of iGDB's potential to facilitate future research.

4.1 Identifying AS spatial extent

Many research questions require an understanding of the geographic locations of logical Internet infrastructure, such as a network's peering locations. Having standardized access to this information allows researchers to focus on their area of study, without having to generate the information on their own. We provide two examples here of how iGDB enables an understanding of AS spatial extent.

The database relations of iGDB allow us to easily identify the geographic overlap of ASes. We illustrate this use case by determining the geographic overlap between Cox Communications and Charter Communications, two large residential cable broadband access ISPs in the United States. We first execute a SQL query in iGDB to identify the ASNs associated with the two organizations. Cox Communications owns one ASN (22773) and Charter Communications uses four ASNs (20115, 7843, 20001, and 10796). We then execute additional SQL queries to identify the unique urban areas in which they advertise a peering presence as well as the locations where they overlap.

Figure 6 depicts the spatial extent and overlap of both ISPs with each metro area in which they operate annotated by a circle. Cox Communications has a presence in 30 urban areas and Charter Communications has a presence in 71 urban areas, with an overlapping presence in only 10 urban areas (Alexandria, VA; Atlanta, GA; Chicago, IL; Cleveland, OH; Dallas, TX; Irvine, TX; Los Angeles, CA; New York, NY; San Diego, CA; and San Jose, CA). This functionality of iGDB allows researchers to easily compare where network operators are making investments in network infrastructure.

Table 2: ASes with physical presence in the most countries.

ASNumber	ASName	Organization	Countries
13335	CLOUDFLARENET	Cloudflare, Inc.	52
6939	HURRICANE	Hurricane Electric LLC	50
8075	MICROSOFT-CORP	Microsoft Corporation	50
174	COGENT-174	Cogent Communications	45
16509	AMAZON-02	Amazon.com, Inc.	44
42473	AS-ANEXIA	ANEXIA Internetdienstleistungs GmbH	44
32934	FACEBOOK	Facebook, Inc.	42
32261	SUBSPACE	SUBSPACE	41
20940	AKAMAI-ASN1	Akamai International B.V.	38
15169	GOOGLE	Google LLC	35
57463	NetIX	NetIX Communications JSC	35

Alternatively, combining data from multiple tables in iGDB enables a researcher to quickly sift through information from multiple data sources with attributes describing over 25k ASes to determine those with physical peering presence in the most countries, as seen in Table 2.

Areas of study enabled by iGDB: Having this information readily available allows a researcher to better understand where organizations focus their investment in infrastructure and can help operators identify opportunities to extend the reach of their services.

4.2 Generating physical paths from logical measurements

Tools such as traceroute are often used to identify the path between end hosts. However, due to MPLS tunneling, nodes that appear directly connected at the IP layer may be separated by additional nodes hidden by MPLS. Integrated with other tools, iGDB enables the efficient fusing of logical information (such as IP addresses resulting from intermediate hops along a traceroute) with physical information about the Internet, such as PoP locations and inferred fiber optic cable links, to provide a natural framework to recover geographical paths taken by tunnels.

Although the IP layer was developed to provide end host to end host addressing and routing, it suffers from inflexibility with traffic engineering and route selection as well as speed of processing IP headers. Layer 2 routing using MPLS is frequently used to address these issues [53]. However, the intermediate nodes and paths that packets travel from source to destination via layer 2 routing is hidden from layer 3 topology measurement techniques such as traceroute. Although techniques exist to identify the presence of MPLS tunnels as well as the number of intermediate nodes traversed, no technique exists to determine the geolocation of intermediate nodes in an MPLS network [73].

We therefore demonstrate how iGDB can be used to infer, visualize, and analyze the physical path that packets travel between two nodes by: (1) using tools external to iGDB to elucidate information on the logical path, (2) leveraging iGDB to fuse the logical path with the corresponding physical path that packets travel along, and (3) leveraging iGDB to visualize and analyze the corresponding inferred physical path and alternate paths that may exist between the nodes. The starting point for this analysis is a list of IP addresses from a traceroute collected between vantage points on the public Internet [19, 66]. Next, we execute an rDNS query to identify the AS ownership of each IP address and to identify the hostname for

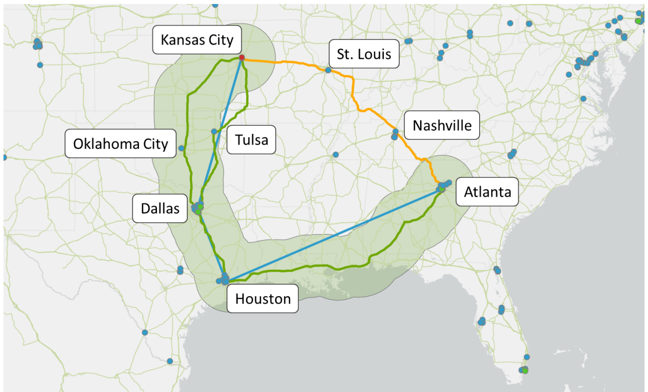


Figure 7: Kansas City to Atlanta traceroute path (blue), inferred physical path through Tulsa, OK or Oklahoma City, OK (green), and shortest practical physical path (orange). AS metro locations depicted as circles with: AS174 (blue), AS12186 (red), AS20473 (green), and AS64199 (purple).

each IP address [13]. With this information, we can often determine the geographic location for each network device.

Previous studies showed that ISPs often encode geohints within the hostname assigned to IP addresses associated with network infrastructure devices [22, 50, 70]. The Hoiho hostname to location geohints are available for use in the form of a set of downloadable regular expressions [4]. For this use case, we determine the city-country code from the hostnames by leveraging these existing regexes derived by Luckie *et. al* for Hoiho [50], rather than learning and developing our own hostname-location pairings.

We next queried iGDB to identify the geographic coordinates associated with each city-country code and to identify all other peering locations for each AS that was traversed along the traceroute. We then plotted the AS peering locations along with straight lines connecting the hops from the traceroute. An example of a traceroute from a RIPE Atlas anchor in Kansas City, MO to an anchor in Atlanta, GA is seen in Figure 7 (blue lines), in which the geolocation of IP addresses identifies the path through Dallas and Houston before reaching Atlanta. We next queried iGDB to identify the inferred physical paths that connected each of the city-country code pairs observed in the traceroute.

By plotting the results of this analysis on a map, we are able to gain two insights, (1) we can identify candidate intermediate nodes between hops that are not observed in the traceroute, for example because of MPLS tunnels, and (2) we can see that the path determined by the intermediate hops in the traceroute is not the shortest geographic path between the start and end nodes.

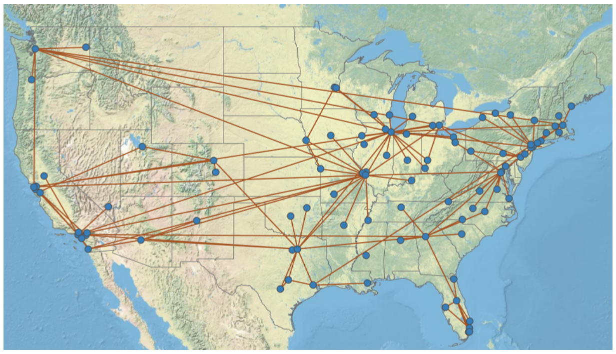
To determine if an intermediate node is physically traversed but logically hidden (*i.e.,* through an MPLS tunnel) we execute the following methodology. We used ArcGIS to create a spatial buffer around each inferred physical route and conducted a spatial join to determine if there is a peering location (node) belonging to the AS(es) identified at either end of the physical link. If there is a physical peering location inside the buffer that also has a physical link in iGDB, we present a more probable inferred physical path as going through that intermediate node, as seen in Figure 7 (green lines), in which this methodology infers that either Oklahoma City or Tulsa is an intermediate node that is hidden from traceroute.

We next aim to quantify how closely the physical path determined from the analysis above conforms to the *shortest practical physical path* between the source and destination cities in a traceroute. We define *shortest practical physical path* as the geographically shortest route along the inferred physical paths that connect peering points along the long-haul links that connect metropolitan areas. To do this, we first define a network where nodes are the city-country pairs in which any AS has a peering location in iGDB and the edges are the inferred physical paths in iGDB that connect those cities. Each edge is weighted by its geographic length. We then conduct a shortest-path analysis along this network to determine the shortest geographic path along inferred physical network infrastructure between the source and destination nodes. In this example, the *inferred path* is: Kansas City ⟶ Tulsa ⟶ Dallas ⟶ Houston ⟶ Atlanta with a length of 2,518km. The *shortest practical physical path* is: Kansas City ⟶ St. Louis ⟶ Nashville ⟶ Atlanta with a length of 1,282km. We next calculate the *distance cost* as the physical length of the actual inferred path divided by the physical length of the shortest practical physical path. In this example, the distance cost is $2518 \div 1282 = 1.96$. The most geographically efficient paths will have a *distance cost* of one and the path is less efficient as the *distance cost* increases.

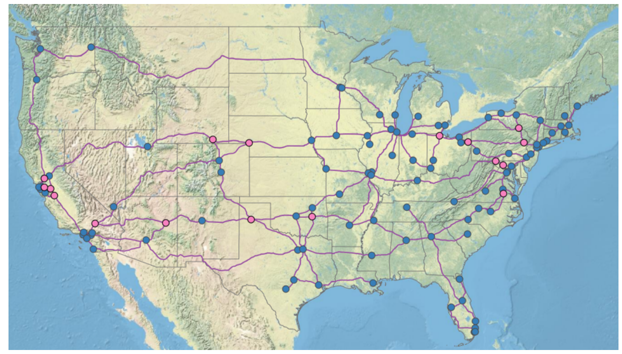
Areas of study enabled by iGDB: Leveraging logical measurements in conjunction with physical topology and shortest-path analysis may be used by network operators to improve network efficiency (*e.g.,* by reducing latency between cities with new long-haul cable infrastructure). This technique could also be used by researchers, who often have access to large traceroute datasets, to identify long-haul cable infrastructure used by ASes of interest at risk from environmental damage (*e.g.,* through a technique like RiskRoute [30]).

4.3 Improving physical representations of logical connectivity

Rocketfuel revolutionized how ISP topologies are analyzed and represented on physical maps by geolocating physical nodes from traceroute measurements and depicting logical connectivity identified through traceroute measurements [74]. These maps, which were validated by ISPs, were compiled in a repository called the Internet Topology Zoo [44] allowing researchers to gain a better appreciation of the geographic extent of ISP networks, including physical nodes and logical connectivity. Those datasets are still widely used more than a decade after as can be seen by the high number (more than 220 papers) of papers referencing the maps in the past 2 years. However, because Rocketfuel depicted connectivity between nodes as straight lines, as seen for AS7018 in Figure 8 (left), researchers could misinterpret the logical diversity of connectivity as indicative of a similar diversity of physical paths. This leads to the question: can we improve the Internet Topology Zoo representation of connectivity between nodes using the right-of-way approach pioneered by Durairajan *et al.* in their study of long-haul fiber-optic infrastructure [29]? The answer, of course, is that iGDB includes both *(1)* a database of information on physical node locations and long-haul fiber optic infrastructure inferred physical



(a) Original Rocketfuel representation



(b) iGDB representation

Figure 8: (Left) Recreation of Figure 7 from [74] showing nodes (blue circles) and edges (brown lines) representing logical connectivity of AS7018 (AT&T) developed with Rocketfuel. (Right) Physical nodes (blue circles), inferred physical nodes (pink circles) and inferred physical paths (purple lines) from iGDB as an alternate depiction of physical connectivity of the Rocketfuel AS7018 map.

paths, and *(2)* a methodology to determine physical paths from logical measurements, as described in Section 4.2.

Starting from the edges between metropolitan areas identified by Rocketfuel for AT&T (AS7018), we first standardize the nomenclature of the metropolitan areas by conducting a spatial join with the cities from the Voronoi diagram in iGDB. From the standard cities, we identify all the inferred physical paths that correspond to the edges validated in the Rocketfuel study through an SQL query in iGDB. Figure 8 (right) shows the results using ArcGIS and we immediately see less diversity of physical paths than what is implied by the Rocketfuel representations. We see that the implied diversity of paths from central California to the east actually proceed along a single physical path through Sacramento and Salt Lake City. In southern California, the physical paths proceed eastward along only three corridors: Las Vegas, San Bernardino to Phoenix, and San Diego to Phoenix. Turning to Florida, we see much less diversity in north-south physical infrastructure than is implied by the Rocketfuel maps.

iGDB includes additional physical paths between other nodes, but our representation of the AT&T network was built from the same node-edge pairings as the original Rocketfuel map. Our map demonstrates real-world constraints to physical paths not visualized with the Rocketfuel representation. Additional physical paths may exist between the AT&T nodes. Insider information on ISP networks may provide specific insight on physical path diversity.

Areas of study enabled by iGDB: By incorporating physical right-of-way constraints, iGDB improves the representation of edges that connect physical Internet PoPs and IXPs, thereby improving our representation of Internet connectivity. Because these paths are generated at low computational cost, they can be used in a variety of research settings, including: risk mitigation for natural disasters (*e.g.,* such as described by Eriksson et. al in [30]), understanding environmental constraints on Internet connectivity, and improving understanding of geopolitical influences on Internet connectivity [69].

4.4 Inferring geographic information from logical measurements

The information in iGDB is gathered from a diverse set of academic and operational sources which serve different purposes: to understand the Internet through observation, measurement, and analysis; and to make the Internet the most effective communication tool for economic, government, education, and other purposes. However, each source of information has drawbacks and incomplete perspective: *(1)* academic sources may be limited in scope when relying on publicly available information, information that can be inferred or measured, or limited access to pockets of proprietary information; *(2)* operational sources may overstate capabilities or share limited proprietary information with the public to achieve their business objectives; *(3)* ASes might not actually own any of the equipment at the interconnections due to remote peering and other gateway technologies that permit an AS to have its traffic carried by a third party leading to a virtual presence in PoPs. Because of these drawbacks, we hope that the community will use inference techniques to extend the data available in iGDB. To demonstrate how these inferences could improve the data available in iGDB, we showcase one example here: inferring the presence of an AS in a metropolitan location that is not present in the baseline iGDB.

We use a simple approach inspired from *belief propagation* in Bayesian networks [54, 55] that uses iGDB to infer the metropolitan locations where traffic is transferred between two ASes. As a starting point, we gather a large corpus of traceroute measurements from RIPE Atlas [66]. We collect AS ownership of each IP address found in the traceroute measurements through bulk queries to bdrmapIT [51]. We next connect the logical IP addresses to the physical locations in iGDB through an rDNS geolocation technique including Hoiho [50] and IXP IP address [56]. When conducting hostname geolocation using Hoiho, we use the existing Hoiho regular expressions and do not learn or generate new ones. Note, however, that for some IP addresses, we were either unable to resolve the address to a hostname during the rDNS lookup (36% of all the observed IP addresses) or that the IP's hostname did not

Table 3: Missing locations in Internet Atlas and PeeringDB for AS174 (Cogent Communications)

Reverse Hostname	Metro
*be2695.rcr21.**drs**01.atlas.cogentco.com*	Dresden-DE
*be3172.rcr21.**syr**01.atlas.cogentco.com.*	Syracuse-US
*be3701.ccr21.**hkg**02.atlas.cogentco.com*	Hong Kong-HK
*be3641.rcr52.**mco**01.atlas.cogentco.com*	Orlando-US
*be4445.287.rcr51.**ktw**01.atlas.cogentco.com*	Katowice-PL
*te0-0-2-3.nr11.b006412-5.**jax**01.atlas.cogentco.com*	Jacksonville-US

contain any geographic hints and could not be geolocated (86% of the entries resolving). While a higher geolocation rate would be better, these provide a sufficient starting point for our example.

We then combine latency measurements from the traceroute data with IP addresses that have known locations to infer the geographic presence of an AS in the following manner. We start with a segment of a traceroute that includes adjacent IP addresses, IP_A—IP_B that correspond to the AS adjacency AS_A—AS_B. In the case that IP_B has been geolocated, but IP_A has not been geolocated, the goal is to infer a geolocation for IP_A, and therefore for AS_A. If the observed differential latency between IP_A and IP_B is less than 2 ms and both IP_A and IP_B are within 30 ms of the host that initiated the traceroute, we infer that IP_A is in the same location as IP_B. We choose 2 ms as the boundary between metropolitan locations for those reasons described in [25, 54, 57], but different latency boundaries could be chosen to be more or less restrictive depending on research objectives.

We conduct a round of inferences on all IP addresses from the original traceroute measurements (*i.e.,* RIPE Anchor meshes). This results in an expanded set of IP addresses (and ASes) with geolocation and therefore a new set of adjacent IP addresses in which one IP address has a known geolocation and the other is unknown. Because of this, we repeat these inferences in a series of iterations, with each iteration including an expanded set of geolocated IP addresses.

Both rDNS geolocation and latency geolocation are established techniques to determine the physical location of logical descriptors (IP addresses and ASes). We showcase their utility here to extend the information in iGDB *because* they are well-accepted techniques. We clearly tag each inference in iGDB so that users may discard the inferences if desired for their research or operational purposes.

To illustrate the utility of these techniques to improve our understanding of physical topology, we observe that more than 80% of the locations identified through reverse DNS do not appear in the initial version of iGDB. Furthermore, reverse DNS provides geographic information on 177 ASes with no known geographic locations. We take AS174 (Cogent Communications) to illustrate how each inference technique can help to uncover missing PoPs. In that context, we are able to find more than 104 cities that were missing, a subset of which are shown in Table 3.

To demonstrate the correctness and benefits of our belief propagation technique in our dataset, we first count the number of new entries in *asn_loc* that we are able to add at the end of the process. We obtain 2231 new (city-AS) tuples in more than 124 metros and 240 ASes through a single iteration of our belief propagation. We quantify consistency across both our inferences; for every IP address, we look at *(1)* its inferred geographic location according to our belief propagation, *(2)* its inferred location according to Hoiho or *(3)* its true location according to IXP prefixes. When an IP address possesses both an informative rDNS or is location in an IXP on top of a location through latency constraints, then 86% of the output from belief propagation results in recovering the same metro area. Building automated consistency checks is simplified with iGDB since it enables self-contained SQL queries instead of writing scripts.

Areas of study enabled by iGDB: The bulk of Internet traffic is routed through peers at private facilities which are less likely to be documented in public datasets. Furthermore, we anticipate that as geopolitical pressures on ISPs increase (see *e.g.,* [23, 31, 59, 69] for concrete examples of governmental pressures in the past), the accessibility and incentive to share information on PoP locations publicly will decrease. Improving the accuracy of information related to AS locations will be helpful to researchers and operators in areas of study and operations related to Internet outages, disaster mitigation steps, understanding the effects of peering relationships on business goals, latency aware traffic engineering, load balancing and effects of BGP announcements.

4.5 Real world comparison with a theoretical example

As a final case study, we return to the theoretical example traceroute from a user in Madrid, Spain to Berlin, Germany originally described in Section 1 and depicted in Figure 1. That example showed our original motivation for this study to fuse logical information with physical attributes through *geographic location* and how we envisioned the manifestation of this fusion. But, does iGDB truly provide the data and methodology to realize that vision?

For a real-world starting point, we consider a recent traceroute measurement from a RIPE Atlas anchor in Madrid, Spain to a RIPE Atlas anchor in Berlin, Germany on 3 May 2022. This particular measurement included 11 IP addresses in which every hop along the traceroute responded with a TTL expired ICMP message. We determined the AS ownership and rDNS for each of these IP addresses using the methodology outlined in Section 4.2. The IP addresses belonged to three ASes: AS20647 (IPB Internet Provider in Berlin GmbH), AS22822 (LLNW), or AS12008 (ULTRADNS). Of the three ASes identified in the traceroute, AS20647 is a regional ISP with AS peering locations in Germany, Netherlands, and Belgium; AS12008 has a peering presence in 18 countries (7 European); and AS22822 is the most geographically distributed with a peering presence in 29 countries (17 European). The European peering locations are depicted as circles in Figure 9 and the spatial extent of the European peering locations is shown as the translucent polygons.

We geolocated 7 of the IP addresses with Hoiho directly and the other 4 IP addresses with RIPE geolocation services [20]. The measurement, depicted as a curving brown line in Figure 9, traveled through five cities: Madrid, ES; Paris, FR; Frankfurt, DE; Düsseldorf, DE; and Berlin, DE. There were four hops each in Madrid and Berlin, but a single hop in each of the intermediate cities. We did not identify any additional candidate intermediate hops in cities that were not identified from the IP addresses we geolocated.

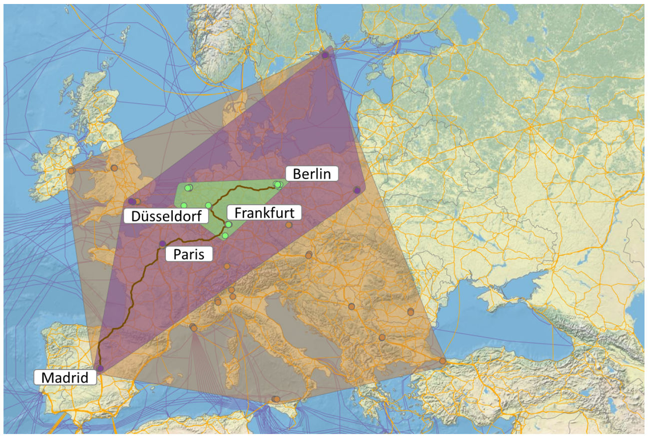


Figure 9: iGDB fusion of logical data anchored by geographic location for a traceroute from a RIPE Atlas anchor in Madrid, Spain to another anchor in Berlin, Germany.

We show how this real-world traceroute measurement compares with the theoretical example from Figure 1 in Figure 9 and quantify the differences here. The IP addresses belong to three ASes, as compared to the four we originally surmised. There are hops in five cities along the path, as compared to 10 in our theoretical example. The physical path goes through only three countries, compared to the original 6 we theorized. The geographic path shown in the real-world traceroute seems a far more reasonable estimate as compared to the straight-line paths between cities in Figure 1. Finally, the AS spatial extent is far more broad than we originally depicted. Each of these differences highlights the importance of iGDB to the research and operational communities to better bridge the gap from logical to physical attributes through the underlying *geographic location*.

5 DISCUSSION

Our system depends on existing research and open databases with subsets of physical and logical information that, when combined, provide a more comprehensive platform to enable further research and improvements to deployed networks. While our database is enabled by external resources, we also suffer from the limitations of each component, as described here.

IP to AS mapping: While bdrmapIT combines two state-of-the-art techniques to map *network borders*, we utilize it in our system in a way that differs somewhat from its original purpose. Specifically, we use bdrmapIT to identify AS paths based on traceroute data. This is conceptually and practically an easier problem than bdrmapIT's objective of inferring the routers at network borders. While AS path identification is a natural result of the border mapping process, we have not fully investigated the quality of the AS path inferences. In future work, we are planning to assess the level to which bdrmapIT provides accurate inference of AS paths from traceroute data.

Declarative information: Most of the side-information used in our system is issued from declarative sources such as PeeringDB, rDNS meta information, AS addresses as declared in the *WHO-IS* registry, and so on. These sources can be out of date and/or incorrectly specified. Unfortunately, there are currently no mechanisms for assuring the freshness or correctness of this data. To mitigate this, we provide multiple sources of information when possible to allow for cross-checking across sources to identify discrepancies. A future direction we are planning to investigate is to create a capability to systematically audit these data sources so that providers might be notified when they are no longer accurate.

Incompleteness: Our current system does not claim to provide complete visibility on all Internet infrastructure at both a physical and logical level. Rather, our system is a resource to benefit the research community in how it combines existing sources of information in a succinct and coherent database to enable further research. In particular, iGDB retains the biases of the input datasets, including lack of definitive information on Internet infrastructure in some geographic regions, incomplete public information on private physical infrastructure, and imperfect knowledge of peering relationships in private peering facilities. Our focus for iGDB is on publicly available information on physical and logical Internet infrastructure. We therefore do not explicitly include information on Content Delivery Networks (CDNs), although this information could be added in the future.

Anycast IP addresses pose an interesting problem when attempting to geolocate them because the same anycast IP address could be served by infrastructure in different geographic locations. In future versions of iGDB, we could easily imagine an extra column added in the IP address table that annotates whether an IP address is part of an anycasted prefix. This allows for several locations to be stored for anycast IP addresses.

As new measurement platforms are developed and deployed, our system will improve in scope and accuracy as we ingest new data. For example, as new techniques are developed to identify the presence of nodes hidden from layer 3 measurement platforms by MPLS and geolocate those nodes, iGDB could be modified with an additional active measurement relation to account for the presence and geolocation of hidden nodes belonging to specific ASes. As another example, with knowledge of the physical location of nodes and connectivity of a microwave network that does not follow transportation rights-of-way, the nodes could be added to the existing *phys_nodes* relation and the physical paths (which would be straight lines from node to node) could be added to the *phys_paths* relation. We would modify the relation with a new column to explicitly annotate the type of link or right-of-way network used to create the link (*e.g.*, microwave, roadway, pipeline, rail, etc.).

We envision that iGDB is extensible, evolvable, and maintainable by the community. Because the Internet is constantly evolving as new infrastructure is deployed, new peering agreements are forged, and new protocols are implemented, we designed the database to enable users to collect new and updated information from a diverse set of sources. We anticipate that each of these sources will have their own information benefits, but also incompleteness in the information that they provide.

6 RELATED WORK

Our work takes support from prior studies on Internet topology measurement, Internet connectivity mapping, and Internet resource geolocation.

Internet topology measurement and characterization: Connectivity at different layers of the protocol stack is intrinsically important to virtually every aspect of Internet behavior and operation. As such, connectivity and topological characteristics at different layers have been the subject of many empirical studies over the years. Prior works have focused on generating maps at a router-level connectivity, see for example [37, 46, 58, 71]). More directly related to our work are prior studies that seek to build maps of PoP-level topologies. Representations of service provider infrastructures at the router and PoP-level were generated through a combination of measurements in [74]. Yoshida *et al.* describe a technique for inferring PoP-level connectivity from delay measurements between residential users [80]. More recently, efforts have been made to develop repositories of PoP-level connectivity based on maps that have been published by service providers themselves *e.g.,* the Internet Topology Zoo [44] and Internet Atlas [27]. Similar to our work, Internet Atlas utilizes GIS, however iGDB combines representations of physical and logical topology and extends previous work with PoP-level *topology generation*. Furthermore, it is not tied to any particular GIS platform and since links are generated, security concerns related to exposing critical infrastructure are reduced.

Mapping Internet connectivity: Many prior studies have addressed the problem of identifying connectivity *within ISPs*. Typically, these have relied on traceroute campaigns to produce accurate maps (*e.g.,* [74, 80]). Another approach is to use search to identify published maps of ISP's physical infrastructure [27, 44]. Our geolocation database is informed by Giotsas *et al.* who used various publicly available data sources and a constrained facility search-based approach to identify infrastructure locations [36]. Similarly, Motamedi *et al.* describe *mi*, a tool for mapping the interconnections within a target colocation facility [54]. While the use case in Section 4.4 is close to the idea described in their paper, our methodology and objective of mapping Internet physical and logical topology is different.

Prior work has also addressed the dual problem *i.e.,* identifying connectivity within geopolitical borders [78**?**]. More recently, some studies have tackled the problem of identifying international frontiers and mapping the geography induced by the topology at the AS-level [24, 45]. We go further by considering the topology and the geographic footprint associated with it. Finally, the objective of our work is similar to prior studies on identifying network borders (*e.g.*[49, 51, 52]) and associating PoPs/co-location centers with physical locations [28].

Internet resource geolocation: Significant efforts have been made over the years to identify Internet resources including IP addresses, routers, and facilities. Techniques for geolocation include reverse DNS lookup [22, 41, 50, 70], IP geolocation [43, 60, 79], and delay-based techniques for geolocating routers and infrastructure [38]. The limitations of these methods are well-known, including when they are used for geolocating infrastructure such as router interface IP addresses [34, 35, 62]. Prior studies have also dealt with the special case of mapping IXP-related public interconnections [15, 57, 64]. Our work benefits from these techniques and we demonstrate their utility to combine their results into a database with logical and physical components.

7 CONCLUSION

Research studies and operational tasks that require maps of Internet connectivity – at either the physical or logical level – appeal to data from a variety of sources. This can often be time-consuming, especially when data from disparate sources is incomplete, in different formats and disconnected. Our objective in this study is the ability to simultaneously analyze and visualize physical and logical attributes of Internet connectivity. To accomplish this goal, we created *iGDB*, which uses *geographic location* to fuse physical and logical maps of connectivity. We implement iGDB in a toolkit that collects updated information from existing data sources, organizes this data into a SQLite database, and generates a PostgreSQL spatial database that enables geospatial analysis and visualization in a geographic information system such as ArcGIS. This capability will enable new and more comprehensive research in areas such as: identifying infrastructure at environmental risk, identifying geographic inefficiencies in logical routing, improving understanding of geopolitical influences on Internet topology, and making inferences about the type and location of physical infrastructure hidden from logical measurements. iGDB is an open resource for the community and we encourage community contributions to maintain and expand its capabilities.

ACKNOWLEDGMENTS

We thank our shepherd, Amreesh Phokeer, and the anonymous reviewers for their invaluable feedback. This work is supported by National Science Foundation (NSF) grants CNS-1705024, CNS-1703592, CNS-2039146, and CNS-2106517. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

REFERENCES

[1]]salamatian [n. d.]. ([n. d.]).
[2] 2006. G591:2006 Telecommunications in Road Reserves - Operational Guidelines for Installations. https://www.commsalliance.com.au/Documents/all/guidelines/g591.
[3] 2014. The Act of 7 May 2010 on supporting the development of telecommunications services and networks. https://digital-strategy.ec.europa.eu/en/library/act-7-may-2010-supporting-development-telecommunications-services-and-networks-poland.
[4] 2021. Hoiho Router Naming Conventions per Suffix Over Time. https://publicdata.caida.org/datasets/supplement/2021-conext-hoiho/web/.
[5] 2021. Natural Earth Populated Places. https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-populated-places/
[6] 2022. ArcGIS. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.
[7] 2022. ArcGIS: Create Thiessen Polygons (Analysis). https://pro.arcgis.com/en/pro-app/2.8/tool-reference/analysis/create-thiessen-polygons.htm.
[8] 2022. eurofiber Fiber-optic network. https://www.eurofiber.com/fiber-optic-network.
[9] 2022. Infrapedia. https://www.infrapedia.com/.
[10] 2022. ITU Interactive Transmission Network Maps. https://www.itu.int/en/ITU-D/Technology/Pages/InteractiveTransmissionMaps.aspx.
[11] 2022. Lumen global network. https://www.lumen.com/en-us/resources/network-maps.html.
[12] 2022. Lumen global network. https://www.dqecom.com/network-map/.
[13] 2022. Team Cymru. https://team-cymru.com/.
[14] 2022. United States Code of Federal Regulations, Title 23 Highways, Chapter I Federal Highway Administration, Subchapter G Engineering and Traffic Operations, Part 645 - Utilities. https://www.ecfr.gov/current/title-23/chapter-I/subchapter-G/part-645.
[15] Brice Augustin, Balachander Krishnamurthy, and Walter Willinger. 2009. IXPs: mapped?. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. 336–349.

[16] Abdelhadi Azzouni, Raouf Boutaba, Nguyen Thi Mai Trang, and Guy Pujolle. 2018. sOFTDP: Secure and efficient OpenFlow topology discovery protocol. In *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*.

[17] V. Bajpai, S. Eravuchira, and J. Schönwälder. 2015. Lessons Learned From Using the RIPE Atlas Platform for Measurement Research. *SIGCOMM Comput. Commun. Rev.* 45, 3 (jul 2015).

[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.

[19] CAIDA. 2020. The CAIDA UCSD IPv4 Routed /24 Topology Dataset. https://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[20] Massimo Candela, Enrico Gregori, Valerio Luconi, and Alessio Vecchio. 2019. Using RIPE atlas for geolocating IP infrastructure. *IEEE Access* 7 (2019), 48816–48829.

[21] Ignacio Castro, Juan Camilo Cardona, Sergey Gorinsky, and Pierre Francois. 2014. Remote peering: More peering without internet flattening. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. 185–198.

[22] Joseph Chabarek and Paul Barford. 2013. What's in a name? decoding router interface names. In *Proceedings of the 5th ACM workshop on HotPlanet*. 3–8.

[23] Thomas M Chen. 2011. Governments and the executive 'internet kill switch'. *IEEE Network* 25, 2 (2011), 2–3.

[24] Frederick Douzet, Louis Petiniaud, Loqman Salamatian, Kevin Limonier, Kave Salamatian, and Thibaut Alchus. 2020. *Measuring the Fragmentation of the Internet: The Case of the Border Gateway Protocol (BGP) During the Ukrainian Crisis*. 157–182.

[25] B. Du, M. Candela, B. Huffaker, A. Snoeren, and k. claffy. 2020. RIPE IPmap Active Geolocation: Mechanism and Performance Evaluation. *ACM SIGCOMM Computer Communication Review (CCR)* 50 (Apr 2020).

[26] Ramakrishnan Durairajan, Carol Barford, and Paul Barford. 2018. Lights Out: Climate Change Risk to Internet Infrastructure. In *Proceedings of the Applied Networking Research Workshop*. Association for Computing Machinery.

[27] Ramakrishnan Durairajan, Subhadip Ghosh, Xin Tang, Paul Barford, and Brian Eriksson. 2013. Internet Atlas: A Geographic Database of the Internet. Association for Computing Machinery, New York, NY, USA.

[28] Ramakrishnan Durairajan, Joel Sommers, and Paul Barford. 2014. Layer 1-informed internet topology measurement. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 381–394.

[29] R. Durairajan, J. Sommers, W. Willinger, and P. Barford. 2015. InterTubes: A Study of the US Long-haul Fiber-optic Infrastructure. In *Proceedings of ACM SIGCOMM*.

[30] Brian Eriksson, Ramakrishnan Durairajan, and Paul Barford. 2013. RiskRoute: A Framework for Mitigating Network Outage Threats. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*. Association for Computing Machinery.

[31] Ksenia Ermoshina, Benjamin Loveluck, and Francesca Musiani. 2022. A market of black boxes: The political economy of Internet surveillance and censorship in Russia. *Journal of Information Technology and Politics* 19, 1 (2022), p. 18–33. https://doi.org/10.1080/19331681.2021.1905972

[32] EuroIX. 2020. EuroIX-IXP database. https://ixpdb.euro-ix.net/en/.

[33] B. Fortz, J. Rexford, and M. Thorup. 2002. Traffic engineering with traditional IP routing protocols. *IEEE Communications Magazine* (2002).

[34] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. 2017. A look at router geolocation in public and commercial databases. In *Proceedings of the 2017 Internet Measurement Conference*. 463–469.

[35] P. Gill, Y. Ganjali, B. Wong, and D. Lie. 2010. Dude, Where's that IP?: Circumventing Measurement-based IP Geolocation. In *Proceedings of the USENIX Security Symposium*.

[36] Vasileios Giotsas, Georgios Smaragdakis, Bradley Huffaker, Matthew Luckie, and KC Claffy. 2015. Mapping peering interconnections to a facility. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. 1–13.

[37] R. Govindan and H. Tangmunarunkit. 2000. Heuristics for Internet Map Discovery. In *Proceedings of IEEE INFOCOM*.

[38] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. 2006. Constraint-Based Geolocation of Internet Hosts. *IEEE/ACM Transactions on Networking* 14, 6 (2006).

[39] Tao Hou, Tao Wang, Zhuo Lu, and Yao Liu. 2021. Combating Adversarial Network Topology Inference by Proactive Topology Obfuscation. *IEEE/ACM Transactions on Networking* (2021).

[40] Packet Clearing House. 2020. PCH peering. https://www.pch.net/about/peering.

[41] Bradley Huffaker, Marina Fomenkov, and Kc Claffy. 2014. DRoP: DNS-based router positioning. *ACM SIGCOMM Computer Communication Review* 44, 3 (2014), 5–13.

[42] Hurricane Electric. 2022. Hurricane Electric Internet Exchange Report. https://bgp.he.net/report/exchanges.

[43] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. 2006. Towards IP geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. 71–84.

[44] Simon Knight, Hung X Nguyen, Nickolas Falkner, Rhys Bowden, and Matthew Roughan. 2011. The internet topology zoo. *IEEE Journal on Selected Areas in Communications* 29, 9 (2011), 1765–1775.

[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 *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 184–194.

[46] L. Li, D. Alderson, W. Willinger, and J. Doyle. 2004. A first-principles approach to understanding the internet's router-level topology. In *Proceedings of ACM SIGCOMM*.

[47] S. Liu, Z. Bischof, I. Madan, P. Chan, and F. 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*.

[48] A. Lodhi, N. Larson, A. Dhamdhere, C. Dovrolis, and k. claffy. 2014. Using PeeringDB to Understand the Peering Ecosystem. *ACM SIGCOMM Computer Communication Review (CCR)* 44, 2 (Apr 2014), 21–27.

[49] Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, David Clark, and KC Claffy. 2016. Bdrmap: Inference of borders between IP networks. In *Proceedings of the 2016 Internet Measurement Conference*. 381–396.

[50] Matthew Luckie, Bradley Huffaker, Alexander Marder, Zachary Bischof, Marianne Fletcher, and K Claffy. 2021. *Learning to Extract Geographic Information from Internet Router Hostnames*. Association for Computing Machinery.

[51] Alexander Marder, Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, kc claffy, and Jonathan M Smith. 2018. Pushing the boundaries with bdrmapit: Mapping router ownership at Internet scale. In *Proceedings of the Internet Measurement Conference 2018*. 56–69.

[52] Alexander Marder and Jonathan M Smith. 2016. MAP-IT: Multipass accurate passive inferences from traceroute. In *Proceedings of the 2016 Internet Measurement Conference*. 397–411.

[53] J. McManus, J. Malcolm, M. O'Dell, D. Awduche, and J. Agogbua. 1999. Requirements for Traffic Engineering Over MPLS. RFC 2702. https://www.rfc-editor.org/info/rfc2702

[54] Reza Motamedi, Bahador Yeganeh, Balakrishnan Chandrasekaran, Reza Rejaie, Bruce M Maggs, and Walter Willinger. 2019. On Mapping the Interconnections in Today's Internet. *IEEE/ACM Transactions on Networking* 27, 5 (2019), 2056–2070.

[55] Kevin Murphy, Yair Weiss, and Michael I Jordan. 2013. Loopy belief propagation for approximate inference: An empirical study. *arXiv preprint arXiv:1301.6725* (2013).

[56] George Nomikos and Xenofontas Dimitropoulos. 2016. traIXroute: Detecting IXPs in traceroute paths. In *International Conference on Passive and Active Network Measurement*. Springer, 346–358.

[57] George Nomikos, Vasileios Kotronis, Pavlos Sermpezis, Petros Gigis, Lefteris Manassakis, Christoph Dietzel, Stavros Konstantaras, Xenofontas Dimitropoulos, and Vasileios Giotsas. 2018. O Peer, Where Art Thou? Uncovering Remote Peering Interconnections at IXPs. In *Proceedings of the Internet Measurement Conference 2018*. 265–278.

[58] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang. 2008. In Search of the Elusive Ground Truth: The Internet's AS-level Connectivity Structure. In *Proceedings of ACM SIGMETRICS*.

[59] Ramakrishna Padmanabhan, Arturo Filastò, Maria Xynou, Ram Sundara Raman, Kennedy Middleton, Mingwei Zhang, Doug Madory, Molly Roberts, and Alberto Dainotti. 2021. A multi-perspective view of Internet censorship in Myanmar. In *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet*. 27–36.

[60] Venkata N. Padmanabhan and Lakshminarayanan Subramanian. 2001. An Investigation of Geographic Mapping Techniques for Internet Hosts. *SIGCOMM Comput. Commun. Rev.* 31, 4 (Aug. 2001), 173–185. https://doi.org/10.1145/964723.383073

[61] PeeringDB. 2020. The Interconnection Database. https://www.peeringdb.com/.

[62] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review* 41, 2 (2011), 53–56.

[63] Rapid7. 2022. Rapid7 Labs Reverse DNS. https://opendata.rapid7.com/sonar.rdns_v2/.

[64] Philipp Richter, Georgios Smaragdakis, Anja Feldmann, Nikolaos Chatzis, Jan Boettger, and Walter Willinger. 2014. Peering at peerings: On the role of IXP route servers. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 31–44.

[65] RIPE. 2020. Routing Information Service – RIS. https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris..

[66] RIPE NCC. 2020. RIPE Atlas Anchors. https://atlas.ripe.net/about/anchors/.

[67] Matthew Roughan, Walter Willinger, Olaf Maennel, Debbie Perouli, and Randy Bush. 2011. 10 lessons from 10 years of measuring and modeling the internet's autonomous systems. *IEEE Journal on Selected Areas in Communications* 29, 9 (2011), 1810–1821.

[68] Routeviews. 2020. “Routeviews”. http://www.routeviews.org/routeviews.
[69] Loqman Salamatian, Frédérick Douzet, Kavé Salamatian, and Kévin Limonier. 2021. The geopolitics behind the routes data travel: a case study of Iran. *Journal of Cybersecurity* 7, 1 (08 2021). https://doi.org/10.1093/cybsec/tyab018 arXiv:https://academic.oup.com/cybersecurity/article-pdf/7/1/tyab018/39765655/tyab018.pdf tyab018.
[70] Quirin Scheitle, Oliver Gasser, Patrick Sattler, and Georg Carle. 2017. HLOC: Hints-based geolocation leveraging multiple measurement frameworks. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–9.
[71] R. Sherwood and N. Spring. 2006. Touring the Internet in a TCP Sidecar. In *Proceedings of the ACM Internet Measurement Conference*.
[72] A. Singla, B. Chandrasekaran, P. Godfrey, and B. Maggs. 2014. The Internet at the Speed of Light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*. Association for Computing Machinery.
[73] J. Sommers, B. Eriksson, and P. Barford. 2011. On the Prevalence and Characteristics of MPLS Deployments in the Open Internet. In *Proceedings of the ACM Internet Measurement Conference (IMC ’11)*.
[74] Neil Spring, Ratul Mahajan, and David Wetherall. 2002. Measuring ISP topologies with Rocketfuel. *ACM SIGCOMM Computer Communication Review* 32, 4 (2002), 133–145.
[75] Dan Swinhoe. 2022. Multiple fibers cut across France, impacting several cities. https://www.datacenterdynamics.com/en/news/multiple-fibers-cut-across-france-impacting-several-cities/
[76] Telegeography. 2020. Submarine Cable. https://www.submarinecablemap.com/.
[77] Vertica. 2022. Well-Known Text (WKT). https://www.vertica.com/docs/9.2.x/HTML/Content/Authoring/AnalyzingData/Geospatial/Spatial_Definitions/WellknownTextWKT.htm.
[78] Matthias Wählisch, Thomas C Schmidt, Markus de Brün, and Thomas Häberlen. 2012. Exposing a nation-centric view on the German internet–a change in perspective on AS-level. In *International Conference on Passive and Active Network Measurement*. Springer, 200–210.
[79] B. Wong, I. Stoyanov, and E. Sirer. 2007. Octant: A Comprehensive Framework for the Geolocation of Internet Hosts. In *Proceedings of the USENIX Symposium on Network Systems Design and Implementation*.
[80] Kaoru Yoshida, Yutaka Kikuchi, Masateru Yamamoto, Yoriko Fujii, Ken’ichi Nagami, Ikuo Nakagawa, and Hiroshi Esaki. 2009. Inferring PoP-level ISP topology through end-to-end delay measurement. In *International Conference on Passive and Active Network Measurement*. Springer, 35–44.

APPENDIX

Ethics

This work does not raise any ethical issues.

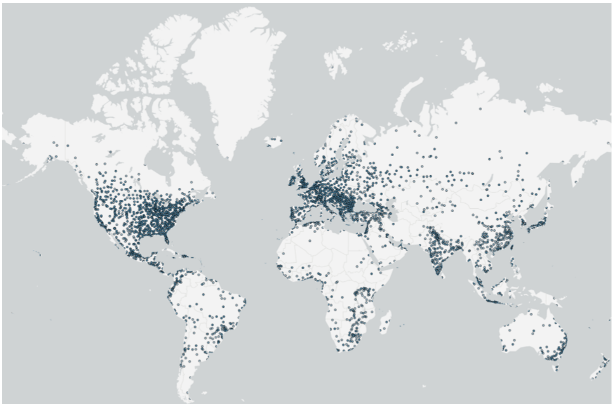
Additional Use Cases

7.1 Geospatial understanding of physical node distribution

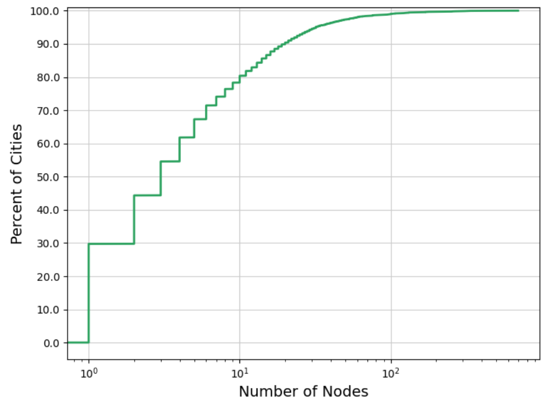
Making use of the layered nature of the data sources in iGDB enables us to conduct geospatial analyses to understand the relationship of Internet infrastructure with other entities in the physical environment. One example is to identify all physical nodes (IXPs, PoPs, etc.) that are in a region of interest, such as an urban area, or to gain a better sense of the geospatial distribution of nodes by attribute, such as all physical nodes operated by an organization of interest. As seen in Figure 10 (left), the physical nodes in iGDB are most dense in central Europe and the eastern United States. We do not claim that this is a fully accurate representation of the physical structure of the Internet, but that the data we gathered from various sources on physical node location, including Internet Atlas and PeeringDB, shows the most nodes in these regions. This also presents an opportunity for community contributions to improve the worldwide representation of the physical Internet in iGDB.

In addition to visual representations of Internet infrastructure, the geospatial nature of iGDB facilitates traditional spatial analyses, such as quantifying the number of nodes located around cities of interest. Although metropolitan areas are a natural geographic division to conduct research on physical Internet infrastructure, they are often represented in spatial datasets as single points. Because the Voronoi division of the Earth’s surface described in Section 3.1 produces a continuous geographic surface with standardized city, state, country labels, it gives us a powerful utility to conduct spatial analysis of Internet infrastructure. To demonstrate the breadth of information in iGDB, we developed an SQL query to determine of the number of physical nodes in each cell of the Voronoi diagram. Of the 7,342 city cells in the Voronoi diagram, 3,130 cells have at least one physical node, with most city cells having fewer than 10 nodes, as can be seen in Figure 10 (right). If desired, we can further select specific cities by country of interest, continent, or by an attribute such as city population to continue the analysis.

Areas of study enabled by iGDB: The knowledge of physical Internet infrastructure in iGDB may be used both to understand its worldwide distribution and may be combined with additional layers of spatial information, such as transportation networks, airports, electricity grids, *etc.* to elucidate the geographic differences in Internet infrastructure deployment worldwide. Augmenting the datasets in iGDB could provide better maps of physical Internet infrastructure, better understanding of deployment choices, or could be used to better understand environmental risks to existing deployments. For example, incorporating additional spatial datasets on powerline and railroad rights-of-way could be used to augment the long-haul infrastructure maps in iGDB; datasets on natural disaster risk, percent of renewable electricity generation, or population demographics could provide insight on datacenter deployments; combining iGDB with datasets on marine traffic, ocean currents, or sea-floor composition could improve understanding of risks to cable landing points.



(a) Physical nodes map



(b) Physical nodes CDF

Figure 10: (Left) Distribution of physical nodes in iGDB shows a concentration of nodes in central Europe and the eastern US. (Right) CDF showing the distribution of number of physical nodes in the Voronoi cell for each of the 3,130 urban areas with at least one physical node.