A Comparative Look into Public IXP Datasets

Rowan Klöti Bernhard Ager Vasileios Kotronis George Nomikos Xenofontas Dimitropoulos
ETH Zurich, Switzerland
{rkloeti,bager,vkotroni}@tik.ee.ethz.ch
FORTH, Greece
{gnomikos,fontas}@ics.forth.gr

ABSTRACT

Internet eXchange Points (IXPs) are core components of the Internet infrastructure where Internet Service Providers (ISPs) meet and exchange traffic. During the last few years, the number and size of IXPs have increased rapidly, driving the flattening and shortening of Internet paths. However, understanding the present status of the IXP ecosystem and its potential role in shaping the future Internet requires rigorous data about IXPs, their presence, status, participants, etc. In this work, we do the first cross-comparison of three well-known publicly available IXP databases, namely of PeeringDB, Euro-IX, and PCH. A key challenge we address is linking IXP identifiers across databases maintained by different organizations. We find different *AS-centric* versus *IXP-centric* views provided by the databases as a result of their data collection approaches. In addition, we highlight differences and similarities w.r.t. IXP participants, geographical coverage, and co-location facilities. As a side-product of our linkage heuristics, we make publicly available the union of the three databases, which includes 40.2 % more IXPs and 66.3 % more IXP participants than the commonly-used PeeringDB. We also publish our analysis code to foster reproducibility of our experiments and shed preliminary insights into the accuracy of the union dataset.

Keywords

Internet Exchange Points, Peering databases, Internet topology

Categories and Subject Descriptors

C.2 [**Computer Systems Organization**]: Network Operations

1. INTRODUCTION

A large part of the interconnection between Autonomous Systems (ASes) in the Internet is realized via *Internet eXchange Points* (IXPs), giving them a major role in the evolution and performance of the Internet. Notably, researchers have recently found that *(i)* the Internet topology is flattening due to IXP-traversing paths which bypass the classic transit hierarchy [12–14, 17], *(ii)* more peerings exist in a single large IXP than in previous sets of measurements for the entire Internet [8], and *(iii)* end-to-end delays and path lengths over IXPs are becoming shorter [9]. Furthermore, IXPs have been proposed as cradles for hosting new technologies, such as Software Defined eXchanges (SDX) [15].

However, the merits and artifacts of the available IXP data have not been thoroughly researched yet. This is in sharp contrast with extensive research on mapping the interconnections between ASes using data from various sources, like RouteViews [2], for more than a decade. A commonly-used source of IXP data in scientific studies, e.g., [10, 11, 18], is PeeringDB [7]. However, in addition to PeeringDB, two other publishers maintain public databases about the global IXP ecosystem, namely the *European Internet Exchange Association* (Euro-IX) [3] and *Packet Clearing House* (PCH) [6]. These datasets are contributed and kept up-to-date by different stakeholders, e.g., by the publisher, or through self-reporting by the IXPs and their participants.

In this work, we do the first cross-comparison of the IXP data provided by PeeringDB, Euro-IX, and PCH. We compare in depth several attributes, like IXPs' locations, facilities and participant information. We highlight the similarity of the available data, complementary information, and data discrepancies. We analyze in total data from about 499, 490, and 687 IXPs in PeeringDB, Euro-IX, and PCH, respectively. To compare the data, we introduce heuristics to link identical IXPs across the three datasets. We find an *IXP-centric* view provided by Euro-IX vs. an *AS-centric* view provided by PeeringDB, reflecting differences in their often volunteer-based data collection approaches.

Besides, we make the linked datasets and our analysis code publicly available [1] to support reproducibility of our experiments and related research efforts. Experiments where this data can be useful include, but are not limited to, *(i)* discovering new peer-to-peer links based on membership data and peering policy so as to augment the Internet topology view, e.g., for modeling the effect of augmented routing protocols [19], *(ii)* investigating the peering ecosystem from a geographical perspective per continent or country, *(iii)* tracking the historic evolution of IXPs and their features, *(iv)* pinpointing the big players in a peering setup, and *(v)* working with new topological paradigms such as IXP multi-graphs [16] in the context of new service provisioning. Compared to using solely PeeringDB, the union of the linked datasets includes data for 40.2 % more active IXPs and 66.3 % more IXP participants.

Finally, we perform a preliminary analysis of the accuracy of the linked datasets and find that even the combined dataset is only 75 % complete when compared to information from BGP route collectors, indicating the need for further research in this context. Partial verification using data available on IXP websites shows more promising results in terms of accuracy, both for the biggest IXPs and for IXPs that are randomly selected from the combined pool of available IXPs. We would like to note though that the three IXP datasets are collected based on voluntary effort and as such, no formal guarantees about completeness, accuracy or freshness can generally be given.

The rest of this paper is structured as follows. We first discuss differences and similarities in particular w.r.t. the collection methodologies of the PeeringDB, Euro-IX, and PCH datasets in Section 2. Then, we introduce our heuristics to link IXPs across datasets in Section 3. We compare the IXP status, location, and facility information in Section 4 and the IXP participant information in Section 5. We discuss and evaluate the accuracy of the datasets in Section 6. Finally, Section 7 concludes our paper and points to future directions.

2. DATA SOURCES

We analyze and cross-compare the three most extensive publicly available IXP datasets, which are provided by PeeringDB [7], Euro-IX [3], and PCH [6]. The datasets inform primarily about IXPs and their participants in varying levels of detail. In Table 1 we compare the types of information and their level of availability in each of the datasets. Importantly, naming and location information is contained in all datasets, enabling us to identify and link identical IXPs in Section 3. We built custom web crawlers and parsers, which we make publicly available [1]. A crawl typically takes between 10 and 30 minutes, depending on the dataset. We acquired all datasets on September 19, 2014. In the remainder of this section, we discuss intrinsic characteristics of each dataset, shedding light on the underlying methodology used by the three data providers to collect and maintain the data.

2.1 PeeringDB

PeeringDB [7] is a worldwide database that aims to serve ISPs which wish to participate in the IXP peering ecosystem. The data available consists of 499 IXPs, their facilities and their participants (i.e., peering ASes). PeeringDB has detailed information about all registered IXPs, unlike Euro-IX which only has detailed information about its affiliate IXPs, while data on non-affiliate IXPs is limited to name, location and status. Moreover, PeeringDB provides detailed information about individual participants, i.e., ASes that peer at IXPs. The data is self-reported by both IXPs and participants.

Notably, some IXPs such as DE-CIX and LINX compile their public membership lists from PeeringDB. This limits the usefulness of these membership lists for verifying the accuracy of PeeringDB.

2.2 Euro-IX

Our second dataset is a list of 490 IXPs provided by the European Internet Exchange Association (Euro-IX) [3]. Its membership consists mostly of European IXPs, which are typically run as cooperative non-profit entities, in contrast to North American Internet exchanges, which are often run as for-profit businesses. Accordingly, European Internet exchanges are generally transparent about peering arrangements. Some of the largest IXPs are located in Europe. Euro-IX supplies information both for affiliated and non-affiliated IXPs. According to the official Euro-IX website [3], "the database information is a combination of both affiliated and non-affiliated IXP content. While the affiliated IXP content is highly accurate, the non-affiliated IXP content is updated on a best effort basis and is nonetheless considered to be quite accurate". From direct communication with Euro-IX staff, we know that the information is generally provided by the IXPs themselves. About two thirds of the IXPs represented have an account to keep their data up-to-date by self-reporting, while 62 of these IXPs (approximately 14 %) have automated the update procedure, which helps improve data completeness and accuracy. Euro-IX provides a website URL and a contact email for all IXPs and participants, i.e., the ASes which connect to an IXP, for 285 of the IXPs. For a subset of IXPs (we assume these are the ones which are registered members of Euro-IX), more detailed information is available (c.f. Table 1). For IXP participants there is limited information, including AS numbers (ASN), name, update time-stamp, IPv6 support capability, and sometimes a URL.

Euro-IX does not provide details about IXPs' individual co-location facilities. However, location information at the city level and, for most IXPs, geographical coordinates are available. Since IXPs can be distributed over several co-location facilities, these location values may not accurately reflect the physical IXP location. For instance, *CyrusOne* is a distributed (likely not Euro-IX affiliated) IXP in Arizona and Texas with points of presence in Austin, Dallas, Houston, Phoenix and San Antonio, but appears in the Euro-IX database only at Carrollton, a suburb of Houston, where its corporate headquarters are located. In addition, Euro-IX does not provide information about IP address prefixes assigned to IXPs, which could potentially be used for linking IXPs across databases.

As a side note, Euro-IX also provides Annual Reports, giving an overview of the state of IXPs in Europe every year.

2.3 Packet Clearing House

Packet Clearing House (PCH) is a non-profit research institute concerning itself with Internet routing and traffic exchange, among other areas pertaining to Internet operation and economics. PCH provides an extensive directory of 687 IXPs [6], including many historical ones. According to direct communication with PCH staff, PCH never removes IXPs from the listing, and marks them defunct only after sufficient verification. 70 % of the IXPs listed are compiled by PCH staff, 25 % are contributed by the Internet community and some 5 % are contributed by the IXP operators themselves. PCH peers at many IXPs itself; the BGP information PCH obtains over these peerings is then used together with DNS and IXP website information (in this order) to derive participant lists. PCH also compiles traffic data from MRTG files (for some 185 IXPs); the other data sources do not have automatic traffic information. For 190 subnets (corresponding to nearly as many IXPs) participant data is entered manually. PCH reports on a per IP address basis, not a per participant basis. As such, an ASN can appear multiple times as a member of an IXP. There are also numerous instances of participant entries containing peering IP addresses but no ASNs. In this study, we only consider entries with ASNs, as we have no other consistent basis for matching the participants across all datasets.

2.4 Data Artifacts

During our data pre-processing and analysis, we observed several artifacts (some quite time consuming) in the datasets, which we report here to simplify future researchers' work.

PeeringDB has two sources of information on connectivity between IXPs and ASes. For each IXP, there is a list of participants, including ASNs. However, for every participant, there is also a list of IXPs. These do not necessarily coincide. A quarter of IXPs present in the PeeringDB dataset have differences between the two sources of information, with more ASNs being listed in the participants' IXP list. This is a consequence of the fact that some participants advertise more than one ASN. The difference in terms of number of participants is 5.7 % on average, although typically no more than a handful of entries. Only 0.5 % of ASNs are responsible for this difference. In general, using the latter data source (participants' IXPs) is preferable due to a slightly higher completeness.

The Euro-IX dataset has 20 IXPs whose participants consist partially, and nine whose participants consist entirely of the reserved ASN "0". In these cases, the administrator has apparently neglected to enter an ASN. These participants contribute about 2 % of the participant entries present, and there are no other duplicate entries.

We also note that PCH has 39 IXPs which have multiple participant entries with the same ASN, with 237 ASNs duplicated in total. Many others have no associated ASN reported at all. As noted in Section 2.3, this is a result of the IP address based reporting used by PCH.

3. LINKING IXPS ACROSS DATASETS

In this section we describe our methodology for identifying and linking identical IXPs in different datasets as well as other pre-processing steps that were necessary to sanitize the data. We use the term *mapping* to refer to identical IXPs that have been linked in two

	IXP																	Members											
Data set	Country and city	Continent	Coordinates	Long Name	Common Name	Status (active)	Media Type (Ethernet, etc)	Protocols supported	Website	Contact information	Costs	Establishment date	Membership requirements	AS Number	Network internals	Associated members	# facilities	Detailed facility info	Organization	ASN	IP address (at IXP)	Company Name	Company Website	Protocols supported	Date Last Updated	URLs	Network details	Policy information	Approx # prefixes	TXT Record	Network status
Euro-IX	✓		+	○	✓	✓	○	✓	+	✓	✓	○		○	○	✓	✓		○	✓		✓	+	✓							
PeeringDB	✓	✓		○	✓		✓	✓	✓	✓				○	✓	✓	✓	✓		✓	+	✓	+	+	+	+	+	+	+		
PCH	+			+	+	+	+		✓			○			○	○			○	○	+	○							○	○	✓

Table 1: Comparison of information available from the Euro-IX, PeeringDB, and PCH datasets. Available = ✓, mostly available = +, sometimes available = ○.

datasets. The key challenge is that IXPs lack consistent identifiers across the datasets. There are several cases of IXPs sharing the same name when they are separate entities, and many cases of identical IXPs being represented by different names in the three datasets. An example is 'SIX'—a name that occurs with minor variations 5 times in PeeringDB (i.e., SIX, S-IX, SIX.SK, SIX SI, SIX NO for Seattle-, Stuttgart-, Slovak-, Slovenian-, Stavanger- IXP respectively). In Euro-IX, there are only three variations of 'SIX', two of which do not directly match the ones in PeeringDB, and at least two different IXPs in Euro-IX share the exact name 'SIX'. In addition, for various reasons (i.e., geographically distributed IXPs), some IXPs exist as single entities in one dataset and as multiple entities in the other.

Due to the large number of IXPs in each dataset, linking all IXPs manually is very tedious and time consuming. Unfortunately, a fully automated approach is not desirable, either, as human expertise is necessary to validate possibly ambiguous mappings. For these reasons, we use a hybrid approach, in which we first automatically produce candidate mappings based on custom heuristics and then we manually verify which candidates actually correspond to the same IXP. Our heuristics to generate candidates for mapping exploit IXP naming and location information and are inclusive in their design. In other words we are conservative in ruling out possible mappings, at the cost of additional manual validation effort.

During our analysis we found that IXPs are sometimes presented at different granularity in the different datasets, e.g., at a facility level in one dataset and as a whole in another. Thus we first merge such sibling IXP records into single entities using the same overall approach as with linking IXPs across datasets. We produce mapping candidates for IXPs that share the same name and location. We explored several schemes for transforming names in order to get good mapping candidates between the different datasets. We apply these name transforming schemes one-by-one, on the original name. After each step, we manually check the produced mappings and remove successfully mapped IXPs from the working datasets. All datasets provide name aliases, which we also take into consideration. Moreover, differences in the location naming convention require additional pre-processing.

Overall, we first merge 26 sibling IXP records into 7 IXPs for a total of 471 IXPs in the Euro-IX dataset, 30 siblings into 12 IXPs for a total of 480 IXPs in the PeeringDB dataset, and 47 siblings into 18 IXPs for a total of 657 IXPs in the PCH dataset. We then use the following heuristics to produce candidates (with the results for Euro-IX/PeeringDB, Euro-IX/PCH, PeeringDB/PCH being respectively reported next to each variant):

1. Directly identical names (214 / 184 / 162 mappings)
2. Converting to lower case (16 / 21 / 26 new mappings).
3. Truncating the name at the second word boundary (2 / 15 / 3 new mappings).
4. Truncating the name at the first word boundary (67 / 101 / 76 new mappings).
5. Removing non-word characters (4 / 8 / 8 new mappings).
6. Various combinations of these, and manual matching (the remaining mappings).

Dataset			Active IXPs			
			Size of		Index	
Euro-IX	PeeringDB	PCH	Intersection	Union	Jaccard	Overlap
✓	✓	✓	273	673	40.6%	73.0%
✓	✓		355	566	62.7%	80.5%
✓		✓	303	512	59.2%	81.0%
	✓	✓	288	566	50.9%	77.0%

Table 2: Intersection and union of the IXP sets which are present in different combinations of datasets, as well as similarity indexes for the sets.

We also explored heuristics based on common IXP member information such as ASNs. However, this turned out to be insufficient in practice due to incomplete reporting of IXP member ASNs (cf. Section 5.1). Another possible attribute that could be explored for linking is assigned IXPs' IP address prefixes. This data is provided by PeeringDB and PCH, but not by Euro-IX. We therefore did not consider it.

In total we find 380, 379 and 344 mappings, respectively. Table 2 shows the size of the intersection (the IXPs that match based on the previous process) and the union (all IXPs) of the datasets, as well as the Jaccard index and overlap index between two sets A and B defined as: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ and $O(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$. Intuitively, the Jaccard index indicates the similarity between sets, while the overlap index indicates the degree to which the smaller set is a subset of the larger. We include both in order to indicate the extent to which the difference is simply the result of one dataset being more complete than the other, rather than the datasets being partially orthogonal. For comparing all three sets we use straightforward extensions of the Jaccard and overlap indices, using all three sets as parameters. All mappings have been manually verified and our approach to generate candidates for mapping is inclusive as explained beforehand. We therefore do not expect false mappings, but we could have missed few mappings in cases we had insufficient or ambiguous information.

We highlight that the datasets provide a lot of complementary information. We interpret this, as well as the differences in IXP names, as indicators that the datasets do not in general have a common source. We further elaborate on this finding in the next section. In total, we find 441, 480 and 374 active IXPs in the Euro-IX, PeeringDB and PCH datasets (after merging), respectively. If we

Location			Number of IXPs		
Continent	Country	City	Euro-IX	PeeringDB	PCH
Africa	*Total*		*31*	*25*	*30*
Asia Pacific	Japan	Tokyo	9	6	11
		Total	*17*	*14*	*23*
	Indonesia	Jakarta	4	8	9
		Total	*6*	*13*	*16*
	Total		*75*	*88*	*116*
Australia	*Total*		*16*	*20*	*23*
Europe	Russian Federation		24	24	19
	France	Paris	9	8	14
		Total	*19*	*20*	*28*
	Germany		16	16	25
	United Kingdom	London	7	12	10
		Total	*15*	*12*	*22*
	Sweden		13	11	14
	Poland		11	12	10
	Total		*201*	*196*	*200*
Middle East	*Total*		*8*	*8*	*10*
North America	United States of America	New York	8	7	14
		Los Angeles	5	3	10
		Chicago	4	4	9
		Total	*92*	*89*	*156*
	Canada		13	16	17
	Total		*110*	*107*	*179*
South America	Brazil		28	41	36
	Total		*48*	*55*	*64*
World	*Total*		*490*	*499*	*687*

Table 3: IXPs in each database by continent. For each continent, we display the countries and cities with the most IXPs. The values reported are based on raw data *before* merging sibling IXPs because some IXPs are distributed in multiple cities.

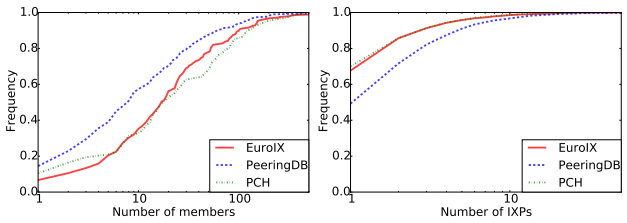
also consider inactive IXPs (e.g., IXPs marked as "defunct" or "unknown") there are 471, 480 and 657 IXPs in the Euro-IX, PeeringDB and PCH datasets. Note that 43.1 % of the IXPs present in the PCH dataset are inactive. We make the compiled datasets available in [1]. Compared to the commonly-used PeeringDB, the combined dataset includes information for 40.2 % more active IXPs.

4. STATUS, LOCATIONS, AND FACILITIES

In this section we compare the PeeringDB, Euro-IX, and PCH databases with respect to the geographical distribution of IXPs, the co-location facilities that house IXPs, and the IXP status information.

4.1 Geographical distribution

All of the datasets contain information concerning the location of IXPs. Based on this, in Table 3 we show the geographical distribution of the IXPs across the globe, and compare how different regions are represented in each dataset. We observe that *the geographical coverage of Euro-IX and PeeringDB is similar, while PCH has somewhat richer coverage in terms of sheer IXP numbers* (including inactive IXPs). On the continent level, Europe has the largest share of IXPs, which corresponds to approximately 40 % in the Euro-IX and PeeringDB datasets and 30 % in the PCH dataset. Interestingly, Euro-IX does not have substantially more IXPs represented in Europe than the other datasets. The next largest region is North America, where PCH has much greater numbers than the other datasets—as discussed in Section 4.3, this is largely due to inactive IXPs. PCH also has a greater number of IXPs for the Asia-Pacific region, with Euro-IX having the least. The other regions are broadly similar. The ranking of the largest countries is also similar across the datasets. The largest cities differ more, with only major world cities being consistently at the top of all of the datasets. In line with our expectations, it appears that more affluent regions have a better coverage by IXPs.



(a) By IXP (b) By ASN

Figure 1: CDFs of the ASes per IXP (Fig. 1a) versus the IXPs per AS (Fig. 1b), for each of the databases. In Fig. 1a, IXPs with no participants are omitted.

4.2 Facilities

Euro-IX provides only the number of facilities for a limited subset of 106 IXPs (22 %), with these IXPs having a mean and median of 6 and 3 facilities, respectively. PCH generally does not provide any facility-related information, although occasionally multiple addresses are listed. In contrast, PeeringDB contains detailed information about facilities, representing them with separate database entities. There are 1,465 facilities listed, 365 of which are in the United States, 126 in Germany, 114 in the United Kingdom, 94 in France and 86 in the Netherlands. The majority of the facilities *are not* associated with an IXP, while 298 IXPs do not report their facilities. 16 facilities are associated with neither IXP nor ISP entities. These observations suggest that *the information on the IXPs' facilities is limited*. Besides, 133 of the facilities associated with an IXP have more than one IXP present, while 112 IXPs are present at more than one facility and 13 are present at more than 10. This indicates that large IXPs are in reality geographically distributed entities. Understanding the drivers and implications of this expansion and transformation that some of these distributed IXPs undergo is an interesting subject for future work.

4.3 IXP status information

The Euro-IX and PCH datasets contain information about the status of IXPs, i.e., whether or not they are currently active. Of all the IXPs in the Euro-IX dataset, 460 are marked as active, 23 as defunct and 7 as under construction. The PCH dataset contains 392 marked active, 90 defunct, 43 planned, 6 deprecated, while 92 have an unknown status. In the PCH dataset 52 entries have the status "not an exchange". Of the 379 common IXPs between these two datasets, 303 share an active status, while 9 share a defunct status. 10 of the matched entries appear as defunct only in the PCH dataset and 4 only in the Euro-IX dataset. Overall, *the status information of the 379 linked IXPs is 82.8 % consistent between the Euro-IX and PCH datasets*.

PeeringDB contains no information on the status of IXPs. Still, a total of 28 PeeringDB entries are marked as defunct in at least one of the Euro-IX (21 entries) or PCH (15 entries) datasets. It is noteworthy that of these 28 IXPs only six report zero participants in PeeringDB, while the others usually report between one and 20, with one IXP reporting 43 participants. We also checked the

websites of IXPs marked as deprecated in Euro-IX or PCH, but yet still reported on PeeringDB. The results showed that most websites cannot be reached or have extremely few members. For example, NWIX Missoula reports only 4 active members, LIX (Luxembourg) has merged with LU-CIX, and five websites don't report an active IXP any more.

Lastly, all but two of the IXPs appearing only in the Euro-IX dataset (38) are marked as active. In contrast, half of the 259 IXPs which are only present in the PCH dataset are either defunct (65) or have unknown status (65), and only 56 of these IXPs are marked as active. Many of the PCH-only IXPs are located in North America. Indeed, according to the PCH dataset, *North America has the largest number of defunct IXPs*, which is likely due to IXPs deployed in the early history of Internet development.

5. IXP PARTICIPANTS

For many use cases, the participants (i.e., peering ASes) of IXPs constitute the most important content of the datasets. Thus, we take a closer look at them in this section.

5.1 IXP-centric versus AS-centric view

Excluding IXPs which have no participants listed, the Euro-IX, PeeringDB and PCH datasets have a mean of 44.3, 27.0 and 30.8 participants per IXP, respectively, with corresponding medians of 17, 8 and 15. This suggests that PeeringDB entries have on average considerably fewer IXP participants listed than Euro-IX entries. Fig. 1a shows the distribution of participant counts for the three datasets. We see that, in general, Euro-IX has the largest number of participants per IXP. Euro-IX provides an *IXP-centric* view as its data is primarily self-reported by IXPs. Besides, IXPs affiliated with Euro-IX typically have a high number of participants—a mean of 104 and a median of 53, contrasting with a mean of 24 and a median of 13 for non-affiliates—as a result of more complete reporting and also because many of the largest IXPs, e.g., LINX, AMS-IX, and DE-CIX, are Euro-IX affiliates. This indicates that large IXPs are generally better represented in the Euro-IX database.

On the other hand, 205 Euro-IX IXPs, 104 PeeringDB IXPs and 636 PCH IXPs have no participants listed. 89 % (53 %) of the Euro-IX (PCH) IXPs which have no participants listed are marked as active. Interestingly enough, seven Euro-IX affiliate IXPs have no participants in the Euro-IX database. Of these, only two separate IXPs appear in each one of the other databases. One of these, CyrusOne, has a limited amount of information about their IXP connectivity available in PeeringDB.

We further analyze IXP participants from the perspective of the participating ASes. The Euro-IX dataset contains records of 6,697 ASes, connected to 1.9 IXPs on average. In PeeringDB, there are 3,784 ASes represented; these are connected to an average of 2.8 IXPs. Finally, PCH contains 1,138 ASes, connected to an average of 1.4 IXPs. 2,167 (Euro-IX), 1,999 (PeeringDB) and 201 (PCH) ASes are connected to more than one IXP; 98, 127 and 5 are connected to more than ten, respectively. Table 4 shows the ASNs which are connected to the largest number of IXPs. We see that Packet Clearing House is among the most prolific peers. PCH's ASN 3856 is used to acquire BGP dumps, reflecting its strategy for data acquisition. PCH's ASN 42 is used for hosting anycasted DNS zones. We also note the presence of large CDNs, like Akamai. Fig. 1b shows the distribution of participant counts from the ASes' perspective for the three databases. The values of IXPs per AS for PeeringDB are generally higher than the values for Euro-IX. These differences likely stem from the mechanisms with which the datasets are formed. In contrast to Euro-IX, PeeringDB provides an *AS-centric* view as its data is self-reported by ASes.

				Number of IXPs		
ASN	Name	Policy	Network Type	Euro-IX	PeeringDB	PCH
20940	Akamai Technologies	Open	Content	61	91	31
6939	Hurricane Electric	Open	NSP	66	84	32
15169	Chief Telecom Inc.	Open	NSP	60	76	24
3856	Packet Clearing House	Open	Educ./Research	50	74	21
42	Packet Clearing House	Open	Educ./Research	44	75	21
8075	Microsoft	Selective	NSP	37	59	22
22822	Limelight Networks	Selective	Content	41	39	18
15133	EdgeCast Networks, Inc.	Open	Content	25	31	18
16509	Chief Telecom Inc.	Open	NSP	21	44	7
10310	Yahoo!	Selective	Content	27	27	14

Table 4: The ASNs connecting to the largest number of IXPs (ranked by the sum). The ancillary information is as reported by PeeringDB.

5.2 Complementarity of IXP participant data

We build IXP-to-ASN *links* for each dataset, which represent $(\mathrm{IXP}, \mathrm{ASN})$ memberships, and perform set-theoretic operations on the extracted links using the Jaccard and overlap indexes as introduced in Section 3. In Table 5 we compare the number and similarity of the IXP participants by continent and IXP sizes.

The Jaccard index of IXP-ASN links between Euro-IX and PeeringDB is at a mere 40 %. Merging PeeringDB with Euro-IX increases the available IXP membership information by 58.9 %. This number goes to 66.3 % when merging PeeringDB both with Euro-IX and PCH. Note that the similarity between the Euro-IX and PeeringDB participant information is greatest in Europe, the region for which both datasets have the largest quantity of membership information (links in Table 5). In the case of Euro-IX, this constitutes well over half of all participant information available. 75 % of the links in Europe (corresponding to 46 % of all links) are contributed by just the Euro-IX affiliated IXPs. Other regions are reported more sparsely, yielding lower similarity: North and South America have Jaccard indexes of 35 % and 38 %, respectively, and other regions have values under 30 %. For the Middle East, the number of participants is so small that the similarity is not meaningful.

As expected, the Jaccard index is much lower for comparisons involving the PCH dataset due to the limited ASN data within the PCH dataset. In terms of the overlap index, the PCH dataset has nearly the same (low) similarity to both of the other datasets, but there are some notable differences between regions: PCH is more in line with Euro-IX within Europe, and otherwise closer to PeeringDB. However, these differences are small in regions with a meaningful amount of information.

Looking at the size categories in Table 5, we find that larger IXPs have a greater similarity, across all pairs of datasets. This holds for both the Jaccard and overlap index.

6. COMPLETENESS OF THE IXP PARTICIPANT DATA

In this section we do a first analysis of the accuracy of the IXP participant information extracted from the three databases. In particular, we try to answer the question of the completeness of the collected information. We cross-compare the collected lists with IXP participant data extracted from 1) live BGP sessions observed in IXP route collector BGP summary data; and 2) 40 IXP websites.

6.1 Comparison with BGP data

In Section 3 and Section 4 we showed that by linking the available IXP datasets we can significantly increase the available information about IXPs and their participants. In this section, we extract IXP participant information from BGP summaries collected by PCH at 77 of their route collectors [5] to compare and evaluate the com-

	Number of links			Euro-IX/PeeringDB		Euro-IX/PCH		PeeringDB/PCH	
Category	Euro-IX	PeeringDB	PCH	Jaccard	Overlap	Jaccard	Overlap	Jaccard	Overlap
Continent									
Africa	247	163	27	23.5%	47.9%	2.24%	22.2%	9.83%	63.0%
Asia Pacific	1049	1105	516	28.4%	45.4%	22.3%	55.2%	22.1%	56.8%
Australia	353	470	49	20.7%	39.9%	6.07%	46.9%	8.81%	85.7%
Europe	7747	5370	1937	46.3%	77.3%	22.6%	92.0%	29.1%	85.1%
Middle East	41	32	27	40.4%	65.6%	47.8%	81.5%	63.9%	85.2%
North America	2059	2436	1009	35.1%	56.8%	25.9%	62.5%	27.2%	73.0%
South America	1088	693	2	38.0%	70.7%	0.0918%	50.0%	0.289%	100%
Size of IXP									
Less than 30	3375	3074	246	24.2%	40.8%	2.52%	36.2%	4.96%	63.8%
30 to 59	1948	1324	277	31.5%	59.2%	11.1%	80.5%	11.4%	59.2%
60 to 119	2837	2159	855	38.9%	64.8%	18.8%	68.3%	24.8%	69.9%
120 to 239	2064	1749	1041	49.1%	71.8%	33.7%	75.1%	41.3%	78.3%
240 or more	2360	1963	1155	74.3%	93.9%	44.1%	93.2%	49.5%	89.4%
Total	12584	10269	3574	40.1%	63.7%	20.5%	77.1%	25.0%	77.4%

Table 5: The number of IXP-to-ASN links by category, and the Jaccard and overlap indexes between each pair of datasets for each category. The categories used are *continent* and *IXP size*—the latter is computed by averaging over all the datasets in order to yield a consistent classification scheme for the three datasets.

pleteness of the participant information in all datasets, including the linked one. The BGP data include information about established sessions with BGP peers over the IXP in contrast to the partially self-reporting origins of the other datasets. Thus, they are a ground truth for BGP peering sessions. PCH tries to openly peer with all other IXP participants. Still, the data may miss participants who do not choose to peer with PCH. We assume that all peer ASes seen by the IXP route collector peer over the IXP fabric. To verify this, we manually scanned the next hop IPs and ASNs within the summary records to determine which ASNs are actually peering at the IXPs by checking for IP addresses from the prefixes assigned to the IXPs. We used BGP data collected on the 19th of Sept 2014, i.e., the same date as the other datasets, and again successfully linked the IXP identifiers of the 77 available PCH BGP route collectors with the IXP identifiers in the other datasets using AS membership and IP address information. The route collectors contain location information in their name (typically an airport code) which we utilized for further verification of the linked identifiers.

In Table 6 we report the number of IXP-to-ASN links by dataset for the 77 IXPs with BGP route collectors and the Jaccard and overlap similarity between the reference BGP data and the four other datasets. First, we find that approximately 72 % of the BGP IXP-to-ASN tuples are reported in the linked dataset, while the corresponding figure is 65.8 % for PeeringDB and lower for the other datasets. Moreover, we find that Euro-IX and PeeringDB include many IXP-to-ASN links which are not present in the BGP data. This indicates that the BGP data is not complete, either. In particular, the route collectors report only approximately 56 % of the membership contained in the databases. The underlying reasons include the fact that not all IXP participants are willing to peer with a route collector, and that the databases may contain stale data.

Besides, the validation dataset used in our study (and in all similar validation studies) is subject to selection bias, i.e., bias due to the IXPs and/or ISPs that provide useful information for validation. Indeed, looking at our set of 77 IXPs we find that the PeeringDB, PCH and Euro-IX datasets are in larger agreement for this validation set than for the overall comparison. For example, PeeringDB and Euro-IX now have a Jaccard similarity of 53.1 % as compared to 40.1 % in the earlier analysis (cf. Table 5). We conclude that the figures presented on dataset completeness in the 77 IXPs may be positively biased. This indicates that the information we have about the completeness of the available IXP participant data, even after linking multiple databases, may be still largely incomplete.

6.2 Comparison with IXP website data

We extracted participant lists from IXPs' websites as an additional source of cross-verification. In particular, we designed custom crawlers for 40 IXP websites providing public membership information in total, which include *(i)* the 20 IXPs with the highest number of participants, and *(ii)* 20 randomly selected IXPs. We selected two sets of IXPs to mitigate the problem of the selection bias we discussed above. If the website of an IXP did not list participant information, then we selected a further IXP either by size or randomly from the two lists above. In total, we had to inspect the 29 largest IXPs to define the top-20 set, and 49 random IXPs for the random set. We collected the website data during the 2nd half of August 2015. At the same time we extracted and linked fresh data from Euro-IX, PeeringDB, and PCH for the selected IXPs to compare fairly with website data.

From IXPs' websites, we extracted in total 6,182 IXP-to-ASN *links* for the top-20 IXPs and 1,181 *links* for the 20 random IXPs. We find that 94 % of the links in the top-20 IXPs are reported in the union of PeeringDB, Euro-IX, and PCH. This number changes to 85 % for the 20 random IXPs. In Fig. 2 we show the common information (i.e., the Jaccard index) between the websites and the linked dataset, and the information only in one of the two sources for each of the top-20 IXPs. We order IXPs by the percentage of common links. We see that for most websites the fraction of common links is above 80 %. For many IXPs, we observe that the linked datasets contain more IXP-to-ASN links than the websites of the IXPs. Only 6% of the links are present only on websites. In contrast, 14 % of the links are present only in the linked dataset. Interestingly, this shows that the union of the three databases suggests a higher number of IXP participants than the websites of the IXPs themselves.

7. CONCLUSIONS AND FUTURE WORK

The quest for representative datasets is perpetual for the research community. Taking into account the rising interest in IXP-related data, in this work we *(i)* compared three rich IXP datasets in order to assess their strengths and weaknesses, and *(ii)* combined them in order to improve the completeness of the publicly available IXP data. Our results show that the three datasets have similar geo-

Number of links					BGP/UNION		BGP/Euro-IX		BGP/PeeringDB		BGP/PCH	
BGP	UNION	Euro-IX	PeeringDB	PCH	Jaccard	Overlap	Jaccard	Overlap	Jaccard	Overlap	Jaccard	Overlap
6,425	8,121	6,087	5,749	3,547	46.1%	71.5%	42.2%	61.0%	45.1%	65.8%	35.3%	73.4%

Table 6: The number of IXP-to-ASN links by dataset for the 77 IXPs with BGP route collectors as well as the Jaccard and overlap indexes. UNION denotes the linked dataset containing PeeringDB, Euro-IX, and PCH.

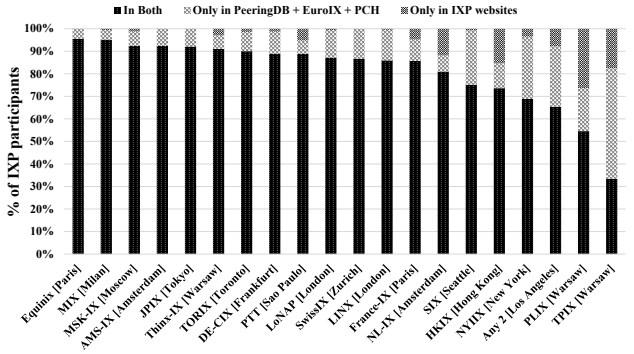


Figure 2: Common and complementary participant information in IXP websites and in the union of PeeringDB, Euro-IX, and PCH datasets. We show the top-20 IXPs with public participant data in their websites.

graphical coverage, with PCH having many more IXPs, including inactive ones. In addition, PeeringDB has an AS-centric bias, while Euro-IX has an IXP-centric bias due the nature of the self-reporting methodologies used by the two providers. Furthermore, our results show that the datasets have partially common as well as rich complementary information. With respect to complementary, we show for example that by linking the datasets we increase the number of IXP records by 40.2 % compared to using solely PeeringDB. Even more complementary information is available for IXP member information, which previous studies have also shown to be incomplete in PeeringDB [18, 20]. Finally, to aid future research, we have made the dataset snapshots as well as the mappings we constructed available to the public, together with the code used to construct them [1].

Still, our results show that while the datasets are partially consistent, they are also incomplete. In particular, the datasets appear to be largely in agreement on the existence of IXPs, and certain attributes such as their operational status. Some of the datasets offer better quantity for certain geographical regions, e.g., Euro-IX for Europe and PeeringDB for the US. However the consistency between the datasets w.r.t. the IXP participants is surprisingly low. We have to stress that it is unclear to which degree these differences stem from under-reporting, resp., from over-reporting such as outdated information. Our study is a first step towards an in-depth analysis of IXP datasets.

The study opens a number of questions for future work. It is currently unclear to which degree data is copied between datasets. While we have not found evidence of copying, it is well possible that data fragments find their way from one database to another, e.g., through third-party contributions. In addition, we would like to understand how the datasets can be cleverly combined, exploiting their individual strengths to improve the accuracy of the available data. In particular, the ground truth behind the available IXP data is still elusive and hard to determine. Other sources of possible ground truth we did not explore in this work are: *(i)* IXPs' looking glass servers, *(ii)* IXPs' newsletters, and *(iii)* event/feeds at IXP websites, which announce new IXP members. A final line of enquiry is understanding the growth trends and consistency of the IXP datasets over time within the evolving Internet peering ecosystem.

We would like to note that PeeringDB and the Internet Exchange Federation (IX-F) started to collaborate, aiming at a more rigorous separation between the roles of the databases as self-reporting points for ISPs and IXPs, respectively. The IX-F acts as a platform for affiliated IXP associations, including Euro-IX and three more continent-level IXP associations. To achieve this they defined the IXPDB [4] and a RESTful API to automatically exchange data between the IXPDB and PeeringDB. We made the methodology described in this paper available to them early to support this effort.

Acknowledgments

We want to thank Euro-IX, PeeringDB and Packet Clearing House for providing free, publicly available sources of information on IXPs. In particular, we want to thank the staff of Euro-IX and Packet Clearing House for providing us information about how data is collected for those datasets. This work has been partly funded by the European Research Council Grant Agreement no. 338402.

8. REFERENCES

[1] Datasets and Software accompanying the paper. https://bitbucket.org/RKloti/a-comparative-look-into-public-ixp-datasets-partially.git.

[2] The Route Views Project. www.routeviews.org.

[3] European Internet Exchange Association. https://www.euro-ix.net/. Datasets collected on: 2014-09-19, at 21:58 CEST.

[4] IXP Database—IX-F Internet eXchange Federation. http://www.ix-f.net/ixp-database.html.

[5] Packet Clearing House (PCH) - Data. https://www.pch.net/resources/data.php.

[6] Packet Clearing House - Internet Exchange Directory. https://prefix.pch.net/applications/ixpdir/. Datasets collected on: 2014-09-19, at 21:58 CEST.

[7] PeeringDB. https://www.peeringdb.com/. Datasets collected on: 2014-09-19, at 11:22 CEST.

[8] AGER, B., CHATZIS, N., FELDMANN, A., SARRAR, N., UHLIG, S., AND WILLINGER, W. Anatomy of a Large European IXP. In *Proc. of ACM SIGCOMM* (2012).

[9] AHMAD, M. Z., AND GUHA, R. Studying the Effect of Internet eXchange Points on Internet Link Delays. In *Proc. of the Spring Simulation Multiconference* (2010).

[10] AUGUSTIN, B., KRISHNAMURTHY, B., AND WILLINGER, W. IXPs: Mapped? In *Proc. of ACM IMC* (2009).

[11] CHATZIS, N., SMARAGDAKIS, G., FELDMANN, A., AND WILLINGER, W. There is More to IXPs Than Meets the Eye. *ACM SIGCOMM CCR 43*, 5 (Nov. 2013).

[12] DHAMDHERE, A., AND DOVROLIS, C. The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh. In *Proc. of ACM CONEXT* (2010).

[13] GILL, P., ARLITT, M., LI, Z., AND MAHANTI, A. The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse? In *Passive and Active Network Measurement*. Springer, 2008, pp. 1–10.

[14] GREGORI, E., IMPROTA, A., LENZINI, L., AND ORSINI, C. The Impact of IXPs on the AS-level Topology Structure of the Internet. *Comput. Commun. 34*, 1 (Jan. 2011).

[15] GUPTA, A., VANBEVER, L., SHAHBAZ, M., DONOVAN, S. P., SCHLINKER, B., FEAMSTER, N., REXFORD, J., SHENKER, S., CLARK, R., AND KATZ-BASSETT, E. SDX: A Software Defined Internet Exchange. In *Proc. of ACM SIGCOMM* (2014).

[16] KOTRONIS, V., DIMITROPOULOS, X., KLÖTI, R., AGER, B., GEORGOPOULOS, P., AND SCHMID, S. Control Exchange Points: Providing QoS-enabled End-to-End Services via SDN-based Inter-domain Routing Orchestration. In *Research Track of the 3rd Open Networking Summit (ONS)* (2014).

[17] LABOVITZ, C., IEKEL-JOHNSON, S., MCPHERSON, D., OBERHEIDE, J., AND JAHANIAN, F. Internet Inter-domain Traffic. *ACM SIGCOMM CCR 41*, 4 (Aug. 2010).

[18] LODHI, A., LARSON, N., DHAMDHERE, A., DOVROLIS, C., AND CLAFFY, K. Using peeringDB to Understand the Peering Ecosystem. *ACM SIGCOMM CCR 44*, 2 (Apr. 2014).

[19] LYCHEV, R., GOLDBERG, S., AND SCHAPIRA, M. BGP Security in Partial Deployment: Is the Juice Worth the Squeeze? In *Proc. of ACM SIGCOMM* (2013).

[20] SNIJDERS, J. PeeringDB Accuracy: Is blind faith reasonable? NANOG 58, 2013.