Improving the Inference of Sibling Autonomous Systems

Zhiyi Chen[0000−0002−5420−9956], Zachary S. Bischof[0000−0001−6419−6303], Cecilia Testart[0000−0002−2993−5059], and Alberto Dainotti[0000−0001−6444−5656]

Georgia Institute of Technology
{zchen798,bischof,ctestart,dainotti}@gatech.edu

Abstract. Correctly mapping Autonomous Systems (ASes) to their owner organizations is critical for connecting AS-level and organization-level research. Unfortunately, constructing an accurate dataset of AS-to-organization mappings is difficult due to a lack of ground truth information. CAIDA AS-to-organization (CA2O), the current state-of-the-art dataset, relies heavily on Whois databases maintained by Regional Internet Registries (RIRs) to infer the AS-to-organization mappings. However, inaccuracies in Whois data can dramatically impact the accuracy of CA2O, particularly for inferences involving ASes owned by the same organization (referred to as sibling ASes).
In this work, we leverage PeeringDB (PDB) as an additional data source to detect potential errors of sibling relations in CA2O. By conducting a meticulous semi-manual investigation, we discover two pitfalls of using Whois data that result in incorrect inferences in CA2O. We then systematically analyze how these pitfalls influence CA2O. We also build an improved dataset on sibling relations, which corrects the mappings of 12.5% of CA2O organizations with sibling ASes (1,028 CA2O organizations, associated with 3,772 ASNs). To make this process reproducible and scalable, we design an automated approach to recreate our manually-built dataset with high fidelity. The approach is able to automatically improve inferences of sibling ASes for each new version of CA2O.

Keywords: AS-to-organization mapping · Sibling ASes · Whois databases.

1 Introduction

Autonomous systems (ASes) are the basic constituent elements of the Internet routing system, managing routing decisions and resources (*i.e.,* IP address prefixes and routers) under a single administrative unit. An AS is uniquely identified by an Autonomous System Number (ASN), which is assigned by a Regional Internet Registry (RIR) as an identifier in the Border Gateway Protocol (BGP). Each AS is typically owned by an individual organization, and one organization may own and operate multiple ASes. ASes owned by the same organization are often referred to as *sibling ASes*. AS-to-organization mappings act as a bridge connecting AS-level and organization-level information. An accurate mapping between ASes and

organizations are crucial in a range of research endeavors: correct identification of AS ownership can offer insights into determining AS business type [28], and also lead to the deduction of AS relationships [23]; accurate lists of sibling ASes under the same organization can help classify events as benign when monitoring for route leaks [20,22] and BGP hijacks [27]; integrating organizations and ASes facilitates studies of organizational BGP behaviors such as IP space utilization [19], censorship evolution [25], and Internet reputation [21,26].

Despite its importance, compiling an accurate AS-to-organization mapping is still an open problem, exacerbated by a significant lack of ground truth. The Whois databases maintained by the five RIRs are the only available authoritative data sources of AS ownership data. However, they are maintained mainly for operational purposes and do not provide a consistent and up-to-date AS-to-organization mapping for registered ASNs (as discussed in § 4).

The state-of-the-art dataset for inferring AS ownership is the CAIDA AS-to-Organization (CA2O) dataset [9]. CA2O leverages information from multiple fields in Whois databases to infer AS ownership and supplements it with manual input. CAIDA has incorporated the CA2O dataset into their ASRank platform [3], a tool commonly utilized by Internet researchers to determine AS ownership, relations, and size. However, the CA2O includes a number of inaccurate inferences of sibling ASes. For example, at the time of writing, AS16509 and AS14618 both belong to *Amazon.com, Inc*, while CA2O does not consider them siblings. Conversely, the owner of AS9426 is *Westpac Bank*, while CA2O maps it to an Australian telecom company *SingTel Optus* and thus it appears to be one of the 63 sibling ASes. Unfortunately, the reasons behind such problems have not been systematically studied.

In this work, we examine in detail the reasons behind the inaccuracies of sibling relations in CA2O and design a methodology to improve the inferences of sibling ASes. We make several contributions: ***(i)*** We start by comparing the mappings in CA2O with the corresponding Whois data and illustrate how CA2O is susceptible to wrong inferences due to inaccurate information in Whois databases in § 3. ***(ii)*** We also inspect PeeringDB (PDB) data and find that it provides an opportunity for addressing the inaccuracies in CA2O: disagreements between CA2O and PDB on sibling relationships serve as hints of potential errors. ***(iii)*** In § 4, we design a pipeline to automatically discover the disagreements and manually conduct a labeling process to investigate the reasons behind the inaccurate mappings. We identify two main pitfalls of Whois data and illustrate how they influence CA2O. ***(iv)*** Based on our analysis and manual efforts, we construct a dataset (called *reference dataset*) correcting 1,028 organizations (involving 3,772 ASes) in CA2O that include inaccurate mappings. The CA2O dataset contains 8,204 organizations that either have sibling ASes (7,573) or have a single AS according to CA2O but have sibling ASes according to PDB (631). We correct relations for 12.5% of them. ***(v)*** To automate the process of improving inferences of sibling ASes, in § 5, we design an automatic approach to reproduce the reference dataset with high fidelity, which is reusable for each new version of CA2O. ***(vi)*** Finally, in § 6, we present a case study of potential BGP hijacking

events and show how our output dataset better identifies siblings and non-sibling related events compared to CA2O. Our improved AS-to-organization mappings provide useful context for examining hijacking events and forensic investigations. Our output dataset is publicly available to the research community[1].

2 Background, related work, and datasets

2.1 Definitions of organizations and siblings

An organization refers to an entity with an established structure for decision-making that involves and links all its subdivisions and groups. In particular, decisions related to the Internet resources that the organization owns and operates (*e.g.,* IP addresses and ASNs) can be coordinated and managed together. Though it is possible that distinct groups within the same organization could operate different ASNs, an organization has the ability to unify and coordinate the operation if preferred. We consider all ASes legally owned and operated by a single organization as sibling ASes.

2.2 Regional Internet Registries and Whois databases

RIRs maintain the authoritative databases related to the assignment of Internet number resources. RIRs are organizations managing the allocation and registration of resources (*i.e.,* IP addresses and AS numbers), which are obtained from the Internet Assigned Numbers Authority (IANA) [7]. Five RIRs are currently serving different regions of the world[2].

Some countries have a National Internet Registry (NIR), which allocates Internet resources to users in the corresponding economy directly from the related RIR's resource pool. When applying for Internet resources, users in those countries have the option to obtain them from either the respective RIR or the NIR. Currently, NIRs only operate in APNIC's and LACNIC's regions, seven in APNIC[3] and two in LACNIC[4]. Another important element in the hierarchy of Internet resource delegation is Local Internet Registries (LIRs), which are the organizations (usually Internet service providers or hosting providers) authorized by RIRs to sub-allocate IP addresses to the end users.

Every RIR and some NIRs maintain Whois databases containing registration information and contact details for each AS and registered organization. In general, Whois databases are organized in objects that have different fields to record AS information, where AS-objects and org-objects are central in the AS-to-organization mapping scenario. Most AS-objects are associated with an org-object with the *orgID* field. However, some ASes do not have the associated org-object, and instead, the actual owner name is stored in the *descr* field. In

[1] https://github.com/InetIntel/Improving-Inference-of-Sibling-ASes
[2] RIPE NCC, ARIN, APNIC, LACNIC, AFRINIC
[3] IDNIC, CNNIC, JPNIC, KRNIC, TWNIC, VNNIC, IRINN
[4] NIC Mexico, NIC.br

addition, RIRs are not responsible for integrating NIRs' Whois databases into their RIR Whois database. The structures of Whois databases vary significantly across RIRs, as they are influenced by the local RIR registration policies. More details of Whois databases for each RIR are summarized in Appendix § A.

2.3 Related work

Cai *et al.* [17] proposed the first work on AS-to-organization mappings. They emphasized the importance of an organization-level view of the AS ecosystem and presented a clustering method to generate an AS-to-organization dataset using Whois data. Their methodology was concerned with three types of Whois records: ASes, organizations, and contacts, where they clustered records using the *orgID*, *phone*, and *e-mail* fields. The authors validated their output dataset using ground truth information from a Tier-1 ISP. In addition, the authors used public documents, routing data, and Whois data to manually create AS-to-organization mappings for nine multi-AS organizations. This dataset was also used for validating their clustering methodology.

In 2012, Cai *et al.* presented a new clustering approach [18] leveraging company subsidiary information from U.S. SEC Form 10-K, which showed few false negatives and false positives compared to their preliminary work, particularly for U.S.-based companies. However, the ISI ANT lab published only one output dataset in 2012 [15] without further updates.

After the above pioneering works, CAIDA developed an inference methodology to map ASes to organizations. Similarly, they created their own objects for ASes, organizations, and contacts. They grouped the objects into families by commonalities in Whois fields. Validated with the same data of Cai's work, CA2O tuned the method and found the following 9 fields were most efficient: *aut.org_id*, *org.admin_c*, *org.tech_c*, *org.phone*, *contact.phone*, *org.org_name*, *aut.admin_c*, *aut.tech_c*, *aut.owner_c* [5]. The CA2O dataset contains two types of objects: AS-objects and org-objects. The inference methodology associates each AS object with an organization object via the *orgID* field.

The CA2O dataset is integrated into the CAIDA ASRank platform, where only the ASes with the same *orgID* are considered to be siblings. CAIDA collects bulk dumps of WHOIS databases 3-4 times per year and produces the CA2O dataset accordingly [2]. In this work, we use the CA2O dataset released in 2022-07-01, which contains 110,764 ASes.

2.4 PeeringDB and other data sources

In addition to Whois databases, several datasets related to the organizational structure of ASes have emerged. PeeringDB (PDB) [4] is a freely available, user-maintained database of networks, where authorized Internet operators can register and update information about their ASes directly. For the purpose of

[5] *aut* refers to autonomous system

facilitating interconnections, PDB also allows Internet Exchange Points, data centers, and other interconnection facilities to maintain information on the site.

We observed that in some cases, information on PDB is more accurate than Whois and CA2O, particularly in instances of acquisitions or mergers. Since it is important for other organizations to have up-to-date information on a network, such as peering policy and contact information, one possible explanation for why Whois may be less accurate in some cases is that there are more barriers to updating records compared to PDB.

One recent example is that *Akamai Technologies* announced the acquisition of *Linode* [13] on March 21st, 2022. The PDB entry for AS63949, previously owned by *Linode*, was changed to *Akamai Technologies* on March 28th, just one week after the acquisition. However, at the time of writing, the Whois information for AS63949 has not been updated, with the latest update recorded in May 2020. Consequently, CA2O still lists the AS' organization as *Linode*.

Unfortunately, PDB has two issues that complicate the task of accurately inferring sibling ASes. First, PDB also contains outdated information. For instance, *KPN* is a Dutch landline and mobile telecommunications company that acquired *EduTel* in 2012 and *Divider B.V.* in 2017. CA2O correctly maps the included ASes as siblings, while PDB still maps AS39309 to *EduTel* and AS47628 to *Divider B.V.* The second issue with PDB is that the coverage of ASes is relatively low; PDB only contains records for 24,367 ASes, covering only about 23% of all currently delegated ASes. Despite these problems, PDB is an extremely valuable source of information, especially given the lack of ground truth.

Another data source is BGP.tools [1], which aggregates AS data from 10 different sources to provide the basic properties of ASes (*e.g.,* URL, business type) and near real-time BGP information. BGP.tools consistently updates URLs by generating possible URLs from the contact information in Whois and checking the correctness manually. Some ASes also self-report website URLs on BGP.tools. In this work, we leverage these website URLs collected by BGP.tools (§ 5).

3 Comparison between Whois and CA2O

Our first step towards understanding the errors in CA2O is to quantitatively compare the similarities and differences between CA2O and the Whois data. Indeed, Whois records are the only source of data of CA2O, other than selected manual updates. We compare the CA2O-mapped organization with the Whois-associated organization for every AS, whose last modified date in Whois is no later than 2022-07-01. For ASes with the *orgID* field in Whois, we use *orgID* as an identifier to compare CA2O and Whois. For the remaining ASes that do not have the *orgID* field and the associated organizations, we compare the organization name in CA2O with the *descr* field in Whois. The results are shown in Table 1.

For APNIC, CA2O makes about 3% different inferences than Whois, while for other RIRs, the ratios of difference are all less than 1%. The discrepancies mainly come from CA2O grouping ASes into families based on the commonalities of Whois fields (*e.g., phone*) [9], where CA2O invents a new *orgID* to relate

Table 1: Results of the comparison between CA2O and Whois

RIRs	APNIC	RIPE	AFRINIC	ARIN	LACNIC
Different inferences	602	112	1	6	108
Candidate ASes	19,880	36,241	2,146	28,391	12,613

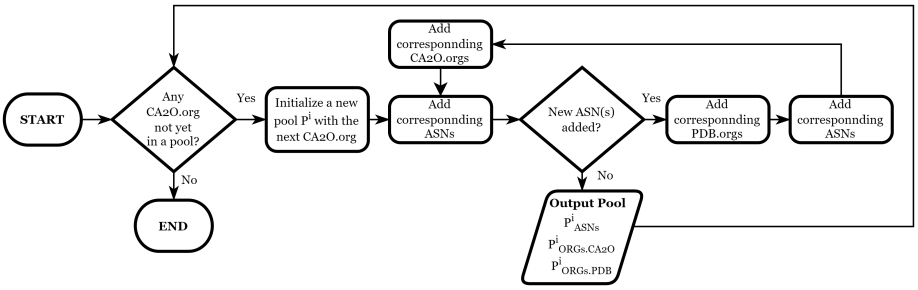

Fig. 1: Detection of *pools* of ASes and organizations that are potentially related according to CA2O and PDB.

the involved ASes (*e.g.,* @family-28933). The results demonstrate that CA2O is so consistent with the Whois mappings that Whois inaccuracies would reflect directly on CA2O (as described in § 4).

To avoid relying on Whois as the single data source, we leverage PeeringDB as an extra dataset. We realize that the disagreements between PDB and CA2O are quite valuable because they help locate potential errors. For example, AS32787 and AS20940 are two famous *Akamai* ASes with big customer cones (*i.e.,* high AS ranks), but CA2O does not regard them as siblings and maps them to different org-objects (*Akamai Technologies, Inc* and *Akamai International B.V*). However, PDB disagrees with CA2O, where the two ASes are siblings under the PDB organization *Akamai Technologies*. The disagreement is a hint directing us to focus on mappings of the *involved* CA2O and PDB organizations. Consequently, we can divide the problem into individual sub-problems based on disagreements, and then conquer them by manually figuring out the real mappings.

Based on the above observations, the roadmap of our work becomes clear: discover all the disagreements between CA2O and PDB, dive into the disagreements to figure out the causes of the inaccurate Whois data (which affect the CA2O dataset), and manually correct the inaccuracies (§ 4). Furthermore, since repeating the manual effort is not scalable, we design an automatic approach, which is able to automatically generate a dataset containing improved inferences of sibling relations for each new version of CA2O (§ 5).

4 Semi-manual investigation

In this section, we dig into the disagreements on sibling relations between CA2O and PDB. In § 4.1, we design a pipeline named *Pool Detection* to automatically

locate and categorize disagreements between CA2O and PDB. To identify sibling relationships and AS-to-organization mappings for each pool, we carry out a manual labeling process as explained in § 4.2. In § 4.3, we identify two pitfalls of the Whois data, which are the causes of inaccuracies in CA2O. In § 4.4, we present the results of our investigation and illustrate how the pitfalls influence the CA2O dataset. Lastly, we briefly introduce a dataset (named reference dataset) produced by our investigation in § 4.6. We refer to the whole effort as a semi-manual investigation because it combines the automatic detection of disagreements and the manual labeling process.

We collected both the CA2O and PDB datasets on 2022-07-01. Our dataset contains 104,153 ASes that were currently allocated by RIRs (i.e., *administratively alive* [24]) according to the delegation files archived on that day.

4.1 Pool Detection

We design the Pool Detection pipeline to automatically discover disagreements between CA2O and PDB in terms of AS sibling relationships. The pipeline groups potentially related ASes and organizations from both the CA2O and PDB datasets into *pools*. For each pool P^i, we use P^i_{ASNs}, $P^i_{ORGs.CA2O}$ and $P^i_{ORGs.PDB}$ to denote the set of AS numbers (ASNs), the set of organizations from CA2O, and the set of organizations from PDB, respectively. Note that it is possible for each pool to have more than one organization from each dataset, due to differences in AS-to-organization mappings between CA2O and PDB. When referring to a specific element in sets $P^i_{ORGs.CA2O}$ and $P^i_{ORGs.PDB}$, we use the notation $CA2O.org$ and $PDB.org$ respectively.

As described in Figure 1, the Pool Detection pipeline examines all organizations in the CA2O dataset in sequence: for each unexamined organization, we initialize a new pool P^i with the organization and CA2O-mapped ASNs; then we start a discovery process to populate $P^i_{ORGs.PDB}$ with PDB-mapped organizations for the set P^i_{ASNs}. We continue the process as long as the PDB organizations include any previously unencountered ASNs. In the end, we obtain pools in which all elements are related, where the ASNs are either (or both) associated with a PDB.org or a CA2O.org. The process is repeated until every organization in the CA2O dataset has been examined.

We next categorize the results of Pool Detection based on the cardinalities of organization sets for each pool. For an output pool P^i, we identify the existence of disagreements by checking if either $|P^i_{ORGs.CA2O}| > 1$ or $|P^i_{ORGs.PDB}| > 1$. For example, if a pool contains more than one CA2O organization ($|P^i_{ORGs.CA2O}| > 1$), there must be some ASes that one of the PDB.org considers as siblings while CA2O maps them to different organizations. In contrast, if $|P^i_{ORGs.CA2O}| = 1$ and $|P^i_{ORGs.PDB}| \leq 1$, it indicates that there is no disagreement on siblings, because neither CA2O nor PDB maps any pair of ASes to different organizations.

The outcome of the Pool Detection process is as follows: initially, we identify 75,041 pools that contain a single AS ($|P^i_{ASNs}| = 1$). The ASes in these pools *do not have any sibling* according to either CA2O or PDB, while the remaining 29,112 ASes in 7,538 pools *have siblings* as per either dataset. Among these, PDB

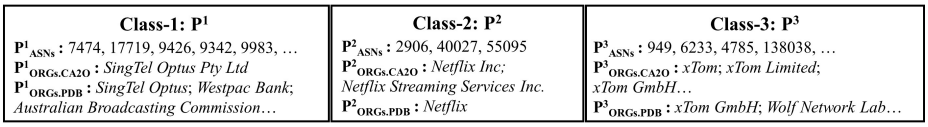

Fig. 2: Real examples of pools for each class.

does not disagree with CA2O on 19,578 ASes in 6,577 pools. There are three possible reasons why a pool may lack disagreement: 1) PDB *completely lacks* any information on all of the ASes in the pool (9,884 ASes in 3,626 pools); 2) PDB *partially agrees* with CA2O when PDB only has information on some of the ASes in a pool (8,588 ASes in 2,475 pools, with PDB having information on 2,923 ASes), or 3) PDB *fully agrees* with CA2O (1,106 ASes in 476 pools).

As the primary objective of this study is to address the disagreements in AS sibling relationships between CA2O and PDB, the remainder of our work centers on pools where PDB and CA2O have conflicting views on AS sibling relationships. This includes 961 pools comprising of 9,534 ASes (32.7% of the total sibling ASes), which are further categorized into the following three mutually exclusive classes based on the properties of each pool:

Class 1 (1:N): $|P^i_{ORGs.CA2O}| = 1$ AND $|P^i_{ORGs.PDB}| > 1$. In this case, the disagreement is that CA2O identifies all the ASes of the pool (two or more) as siblings, while PDB associates them with different organizations.

Class 2 (N:1): $|P^i_{ORGs.CA2O}| > 1$ AND $|P^i_{ORGs.PDB}| = 1$. In this case, the disagreement is that PDB identifies two or more ASes as siblings while CA2O associates them with different organizations.

Class 3 (N:M): $|P^i_{ORGs.CA2O}| > 1$ AND $|P^i_{ORGs.PDB}| > 1$. In this case, the disagreement is due to CA2O finding sibling relationships that PDB does not recognize and vice versa.

Figure 2 provides an illustration of pools belonging to each of the aforementioned classes. Table 2 summarizes the number of pools as well as the number of ASes and organizations for both CA2O and PDB in each class. For *Class-1*, each PDB organization only owns around one AS on average, which indicates PDB recognizes many single-AS or few-AS organizations as different ones from the CA2O organization. The distribution of the number of sibling ASes within CA2O is heavily skewed towards small values (*i.e.,* less than 5) with a long tail, where a total of 973 ASes are identified as siblings under the *DoD Network Information Center*. The distributions of CA2O and PDB in *Class-2* also concentrate on small values. This suggests that PDB recognizes more siblings within some "small" organizations in general, while CA2O identifies them as individual entities. The largest outlier in this class is *VeriSign Global Registry Services* which involves 338 ASes as siblings. *Class-3* shows similar skewed distributions whereas the situation is more complex because each pool contains more than one organization from both CA2O and PDB. We discuss the details of each class in § 4.4.

Table 2: Statistics of the pools with disagreements.

		CA2O		PDB	
Category	**#Pools**	**#ASes**	**#Orgs**	**#ASes**	**#Orgs**
Class 1 (1:N)	544	5,680	544	1,506	1,312
Class 2 (N:1)	337	1,901	791	1,060	337
Class 3 (N:M)	80	1,953	292	817	336
Overall	961	9,534	1,627	3,383	1,985

So far, the Pool Detection locates 961 groups of disagreements where either CA2O or PDB may contain inaccurate mappings. To determine the root cause of inaccuracies and correctly establish AS sibling relationships, we must thoroughly examine each pool individually to identify accurate mappings and sibling relationships. To this end, we perform a manual labeling process in an attempt at obtaining ground truth.

4.2 Manual labeling the pools with disagreements

In the following paragraphs, we introduce the methods we used to identify sibling relationships. We design a manual labeling process: for each pool, we first investigate the relations of every *pair of organizations* (*i.e.,* org-org) to check if they are under the same entity or not, and then we check the correctness of the *ASN-organization mappings* (*i.e.,* ASN-org) for each element in P^i_{ASNs}. It is worth mentioning that we examine all possible pairings of ASN-org, not just the mappings within CA2O or PDB datasets.

We perform four steps to verify the relationships: *(i)* check keywords in Whois names: if organizations contain the same brand name, or if an AS contains the same brand name of an organization; *(ii)* search on Google about relations between organizations (*e.g.,* merger, acquisition, trading name vs. registered name, etc.), or perhaps find the owned ASes on the website of the organization; *(iii)* directly contact operators by email; *(iv)* compare contact roles or persons.

We implement the four-step process for every pairing of elements within a pool (org-org or ASN-org) in sequence and discontinue the process if any resource indicates that two organizations are owned (or not owned) by the same entity or an ASN has an ownership relationship (or no relationship) with an organization.

In a pool, there are three possible outcomes for any two objects:

- Our labeling process confirms two organizations are *under the same entity* or an AS *is owned by* an organization. For instance, by investigating keywords of brand names, we recognize that *Netflix Inc* and *Netflix Streaming Services Inc.* are under the same entity.
- Our labeling process finds evidence that the two organizations are *owned by different entities* or an AS *does not have any relation with* an organization. For example, *Skywolf Technology* (a PDB.org) and *LSHIY Network* (a CA2O.org) are in the same pool, where CA2O maps AS7720 (SKYWOLF-AS-AP) to

LSHIY while PDB maps it to *Skywolf*. By directly contacting *Skywolf*, we confirmed that the two organizations are different, and AS7720 is not owned by *LSHIY* but owned by *Skywolf*.
– All of the four steps fail to find any evidence of an AS sibling relationship, where the brand names are different; the search engine shows no result about the relation; the operator does not reply to our email and the contact information is different. In this case, we consider two organizations are different or an AS is not owned by an organization.

By undertaking the manual labeling process, we gather the mappings and sibling relationships for each pool based on the identified outcomes of pairings, which are expected to be close to the ground truth.

4.3 Two pitfalls of Whois: the causes of inaccuracies

During the manual labeling process, we identify two pitfalls of the Whois data, which are the main causes of the inaccuracies we identified in the CA2O dataset. We verify our findings by consulting the 5 RIRs and some Internet operators. Our paper is the first work that systematically analyzes and characterizes the problems of Whois data across RIRs in the context of AS-to-organization mappings.

APNIC-LIR issue The operation of Local Internet Registries (LIRs) varies among regions due to the diverse policies of the different RIRs. In addition to sub-allocating IP addresses and serving as the upstream of the customer ASes, LIRs under APNIC and RIPE are also responsible for applying for AS numbers on behalf of their customers [14,**?**].

In our analysis, we found that such ASN-related services might cause inaccurate Whois mappings due to the fact that certain LIRs will use their organization identifiers in the *orgID* fields of ASNs obtained on behalf of customer ASes. Consequently, CA2O *incorrectly infers the customer ASes as siblings owned by an LIR*. We consulted with contacts at the five RIRs about such practices and received confirmation that only LIRs in RIPE and APNIC are authorized to provide such ASN-related services. Moreover, only APNIC LIRs associate the AS-objects of customers with themselves: organizations can apply for AS numbers only after becoming APNIC members (*i.e.,* LIRs), while other non-member organizations (*e.g.,* some end users) need to acquire Internet number resources such as ASNs exclusively through an APNIC LIR. In this case, LIRs are responsible for registering AS-objects in the APNIC Whois database for their customers because APNIC considers LIRs as the resource holders for all Internet number resources that they apply for. For example, *LSHIY Network* is an APNIC LIR that only owns 2 ASes and applies for ASNs on the behalf of customer organizations for 26 ASes. In the CA2O dataset, these 28 ASes are all considered siblings under the *LSHIY Network* organization.

Unfortunately, APNIC does not maintain an official list of the APNIC LIRs that provide ASN services, so we need to identify APNIC LIRs ourselves. In

§ 4.4, we demonstrate our manual labeling process successfully identifies APNIC LIRs and discards incorrect inferences of AS sibling relationships in CA2O. In the following sections of our paper, for succinctness, we use the term APNIC LIRs to refer to the LIRs that provide ASN services in the APNIC region, which is technically a subset of all LIRs in the APNIC region.

Multi-orgID issue We define the multi-orgID issue as follows: CA2O splits sibling ASes into different organization objects based on different *orgIDs* in Whois. It is common for all the 5 RIRs to assign different *orgIDs* to groups, divisions, or subsidiaries under the same organization. Since CA2O carries on the Whois information, the CA2O organizations miss sibling relations between these ASes. For example, 7 Amazon ASes are associated with three different *orgIDs* (AMAZON-4, AMAZO-4, AMAZO-139) with the same org-name *Amazon.com, Inc*, where CA2O does not identify these ASes as siblings. It is also possible that one organization owns multiple ASes delegated by different RIRs, so it has to register different org-objects in different Whois databases. Even though the names of organizations are almost the same except for capitalization and punctuation, CA2O infers them as different organizations, such as *University of Guam* in APNIC (AS23676) and *UNIVERSITY OF GUAM* in ARIN (AS395400).

In addition to the above cases, the multi-orgID issue also exists in instances of mergers or acquisitions: the Whois databases may not reflect changes in legal ownership promptly after an acquisition or merger, as the process of updating records can take some time. For example, *GTT* bought *Interoute* in 2018 [6], and AS5580 changed its associated organization in Whois from *Interoute* to *GTT* in 2022. It is also possible that an operator only changes the contact email or auxiliary information (*e.g., remarks* and *descr*) of involved ASes, but does not bother to change the *orgID* and *org-name*. For example, *Agrium* became *Nutrien* by a merger with *PotashCorp* in 2018 [11], and AS137945 added *Nutrien LTD APAC AS* in the *descr* field and changed the contact email, but still kept *Agrium* as the organization name. It is reasonable for operators to do so since updating contact details is enough for operational purposes. Despite this, it is important to address this issue in order to avoid missing sibling ASes in the context of AS-to-organization mappings.

4.4 Results of investigation

In this section, we present the results of our semi-manual investigation and analyze how the two pitfalls influence the inferences of CA2O. The CA2O dataset contains 8,204 organizations either with sibling ASes (7,573) or whose ASes have siblings according to PDB (631), and we *correct relations for 12.5% of them* (1,028 organizations, which are associated with 3,772 ASes). Among the 3,772 mappings of ASes, 580 mappings are impacted by the APNIC-LIR issue and the other 3,192 mappings are impacted by the multi-orgID issue. For the remaining part of the section, we dive into each class (as shown in Figure 3), analyze the influences of the pitfalls, and illustrate with some pools as examples.

Class-1 Our study on *Class-1* reveals that the APNIC-LIR issue is the sole cause of disagreement between CA2O and PDB. In other words, CA2O might wrongly map customer ASes to APNIC LIR organizations but does not miss siblings. Among the 544 pools, we recognize 26 pools that contain APNIC LIRs, where 375 ASes are involved. Our manual labeling process *corrects the mappings of 194* out of 375 ASes by associating the ASes to the actual owners (either PDB.orgs or organizations from *descr*), where we confirm the ownership based on the evidence found by the four-step process above.

As shown in the *Class-1* branch of Figure 3, we first separate the pools whose P^i_{ASNs} contain more than one APNIC-delegated ASes (denoted as candidate APNIC-LIR pools) to locate the possible APNIC-LIRs (remind we do not have an official list of APNIC LIRs), because an APNIC LIR must have at least two APNIC-delegated ASes: one for itself, one for its customer. For the pools impacted by the APNIC LIR issue, CA2O incorrectly maps all *customer* ASes, while PDB is more accurate. Among the 194 mappings that we corrected, 46 ASes have information in PDB, where 42 of them are accurate. For example, *SingTel Optus* is an APNIC LIR, and CA2O considers 63 ASes to be siblings under it. Though PDB only has information for 3 out of 63 ASes, the AS-to-organization mappings are all correct: AS9342 (ABCNET-AS-AP) to *Australian Broadcasting Commission*, AS9426 (WESTPAC-AS-AP) to *Westpac Bank*, AS9438 (NETRO-AS-AP) to *Netro*. Another important observation is that the *descr* field contributes more than PDB when correcting the mappings of customer ASes: 152 out of 194 mappings are corrected based on the *descr*.

The situation is quite different for pools in which CA2O.org is not an APNIC LIR. For the other 64 candidate pools (which we confirm the CA2O.orgs are not APNIC LIRs) as well as the other non-candidate pools, CA2O is very accurate while PDB is not. We identify two problems with the PDB data. First, PDB sometimes over-divides organizations and sibling ASes. For example, PDB wrongly separates *Zettagrid* and *Conexim Australia* as two organizations and breaks the sibling relation between AS7604 (ZETTAGRID-AS) and AS37996 (CONEXIM-NET-AS-AP). Indeed, *Conexim* is a subsidiary of *Zettagrid*, and CA2O correctly identifies the two ASes as siblings. Second, we discover that the PDB information could be outdated. For example, CA2O maps AS21461 and AS44700 as siblings under *Haendle & Korte GmbH* while PDB disagrees and maps them to two organizations (*Haendle & Korte GmbH* and *Transfair-Net*). After consulting the Internet operator by email, we learned that *Haendle & Korte* bought *Transfair-Net*, and AS21461 would be disabled in near future.

To conclude, if the CA2O.org of a *class-1* pool is an APNIC LIR, the mappings of CA2O are problematic for ASes of customer organizations, while PDB is more accurate. In addition, the *descr* field in Whois can be a useful source of information. Otherwise, for the pools without APNIC LIRs, CA2O and Whois are significantly correct, while PDB tends to be inaccurate.

Class-2 For the pools in *Class-2*, the APNIC-LIR issue is unlikely to occur, but the multi-orgID issue often leads to CA2O missing many siblings. Among the

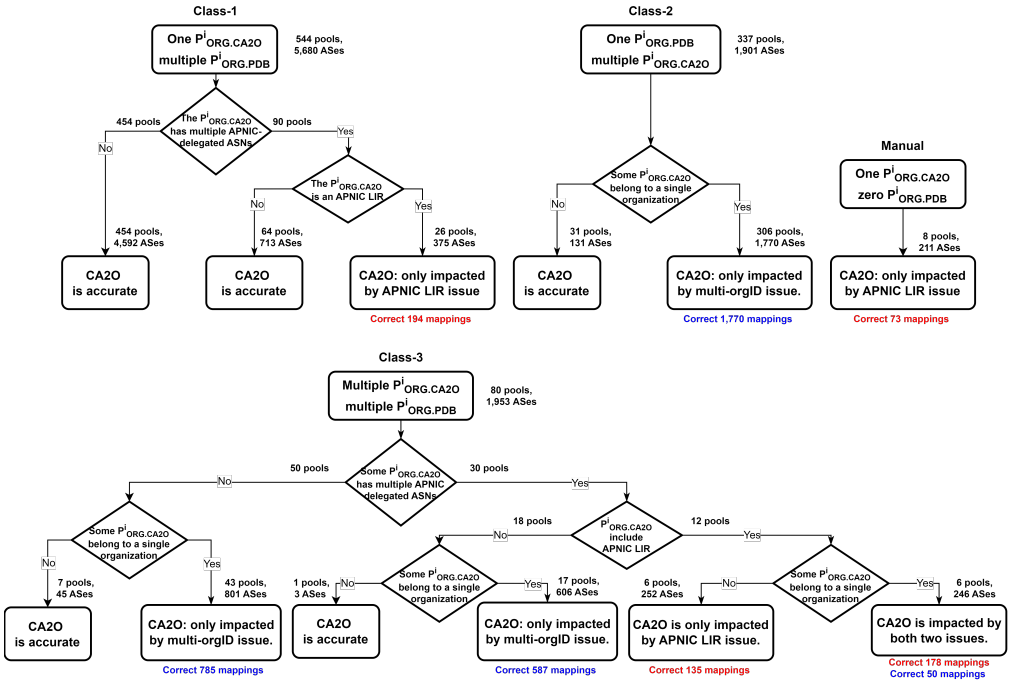

Fig. 3: Results of our semi-manual investigation. The mappings that we corrected are color-coded, with blue indicating those affected by the multi-orgID issue and red indicating those affected by the APNIC LIR issue.

337 pools, we *correct 306 pools (727 CA2O.orgs, 1,770 ASes are involved)* that are impacted by the multi-orgID issue. Our manual labeling process merges the org-objects under the same entity and considers all involved ASes as siblings.

As shown in the *Class-2* branch of Figure 3, PDB is quite accurate for the majority (∼90%) of pools, while CA2O breaks siblings into different org-objects. There are 174 pools where organizations in $P^i_{ORGs.CA2O}$ contain the same brand name in their org-names (*e.g., Netflix Streaming Services* and *Netflix Inc*). In addition, 32 pools are related to acquisitions or mergers (*e.g., Nutrien* and *Agrium*). For the remaining pools, the CA2O.orgs are groups or subsidiaries with different brand names. For example, one of the pools contains two CA2O.orgs (*VIX Route Server*, *ACONET*) and one PDB.org (*University of Vienna*). We directly learned from Vienna University that both *VIX* and *ACONET* are owned and operated by their Computer Center.

For the remaining 10% pools, we consider CA2O to be correct while PDB is problematic because we do not find any evidence to prove the organizations in $P^i_{ORGs.CA2O}$ operate under the same entity. We also observe that the involved organizations usually have different websites or LinkedIn profiles. For example, PDB maps AS24390 (USP-AS-AP) to *AARNet*, while CA2O maps it to *The*

University of the South Pacific. We believe CA2O is correct instead of PDB because *AARNet* is an ISP providing services to the education and research communities in the Australian area.

Although CA2O might miss sibling ASes, the existing mappings within each CA2O.org (*i.e.,* ASN-org) are significantly accurate: 706 out of 791 (∼90%) CA2O.orgs have ASes all with the same keyword in names. For the other 85 organizations, we also manually verify the correctness. For example, AS29697 (CNS) is correctly mapped to *BeeksFX VPS* by CA2O, because we found Beeks group acquired CNS in 2019 [12].

To conclude, the pools of *Class-2* tend to be affected by the multi-orgID issue. For most pools, PDB is correct while CA2O over-divides organizations. However, PDB is not always accurate, as it is problematic in approximately 10% of pools, while CA2O remains correct in those instances. Furthermore, the mappings within each CA2O.org are highly precise, with almost no cases of an ASN being incorrectly assigned to an organization when it actually belongs to another organization.

Class-3 The situation in *Class-3* is more entangled: the two pitfalls might exist simultaneously. Moreover, CA2O and PDB might both be inaccurate in one pool. Our manual labeling process *corrects 223 CA2O.orgs (1,422 ASes included)* that miss siblings, and *corrects 313 mappings of customer ASes* where CA2O maps them to APNIC LIRs.

For 88% of pools without any APNIC LIR (60 out of 68 pools), CA2O misses some siblings due to the multi-orgID issue. In *Class-3*, PDB may sometimes over-divide organizations, unlike in *Class-2* where PDB is always correct on the pools where CA2O is incorrect. We take the pool of *Akamai* as an example: there are 4 CA2O.orgs (*Akamai International B.V.*; *Akamai Technologies, Inc*; *Linode, LLC (APNIC)*; *Linode, LLC (RIPE)*) and 4 PDB.orgs (*Akamai Technologies*; *Asavie Technologies*; *Instart Logic, Inc*; *Nominum, Inc*), where all of the organizations actually belong to *Akamai* because of a series of acquisitions.

In addition to the case where all organizations operate under the same entity, we have observed instances where a pool without APNIC LIRs may contain two or more completely distinct organizations. Indeed, the involved organizations operate some ASes together under partnerships, whereas CA2O and PDB map these ASes differently. For instance, a pool contains 4 CA2O.orgs (*Arabian Internet & Communications*; *Saudi Telecom Company (STC)*; *London Internet Exchange Ltd*; *LINX USA Inc*) and 2 PDB.orgs (*LINX* and *Saudi Telecom Company (STC)*), where all 4 CA2O.orgs miss siblings: the first two CA2O.orgs are actually under the same entity because of an acquisition, and the latter two are subsidiaries. Interestingly, the four organizations fall in the same pool because of AS31177 (JED-IX), which PDB maps to *LINX* and Whois maps to *STC*. In fact, *LINX* and *STC* entered into a partnership to form an Internet Exchange Point called JEDIX in 2018. Our manual labeling process corrects the two organizations by finding the separated siblings and maps AS31177 to *STC* because of the contact email of this AS.

Table 3: A visual example of the reference dataset.

ASN	Reference.orgs	Sibling ASNs	CA2O.org	PDB.org
55095	PDB: Netflix, CA2O: Netflix Inc (ARIN), CA2O: Netflix Streaming Services Inc. (ARIN)	[2906, 40027]	Netflix Inc (ARIN)	Netflix

In the case of pools that contain APNIC LIRs, the main difference from *Class-1* is that some organizations obtain their ASNs from multiple APNIC LIRs, which greatly increases the size of the pools. The biggest pool in *Class-3* contains 14 CA2O.orgs, 97 PDB.orgs, and 155 ASNs. As an example, *YuetAu Network* owns AS147047 and AS138435, which are applied separately through 2 APNIC LIRs (*NEXET LIMITED* and *Aperture Science Limited*).

For the 6 pools that are impacted by both issues, we take the *xTom* pool as an example: *xTom* is a hosting provider which provides services in a wide range of regions. The pool contains 9 subsidiaries of *xTom* delegated by 4 RIRs except for LACNIC (*e.g., xTom Hong Kong Limited*; *xTom GmbH*), where the multi-orgID issue leads to missing sibling relations. Moreover, two of the subsidiaries in the APNIC region are LIRs, where CA2O also makes mistakes on the customer ASes. For example, *xTom Limited (APNIC)* helps *Wolf Network Lab* to apply for AS138038 (WOLFLAB-AS-AP), while CA2O wrongly maps AS138038 to *xTom*.

4.5 Manual input of APNIC LIRs

During our investigation, we identify 8 APNIC LIRs (73 ASes involved), for which we do not find disagreements in their pools, because none of the customer ASes maintain any information in PDB. We manually include them to achieve a more accurate dataset, where details can be found in Appendix § C.

4.6 Reference dataset

By aggregating the results of our semi-manual investigation, we produce a dataset (denoted as reference dataset), which contains our corrections on 1,028 CA2O.orgs (3,772 ASes involved). For the remaining ASes we do not change the mappings, we directly keep the CA2O and PDB mappings in the dataset. There are four columns for each AS number: reference mapping, sibling ASes, CA2O mapping, and PDB mapping, where the first two columns record the results of our investigation. Table 3 illustrates a visual example of the dataset. We welcome corrections from owners of the involved ASes.

Since the Internet world changes rapidly, where new AS numbers could be delegated, and old AS numbers could change ownership, it is not scalable to rerun the semi-manual investigation for every new version of the CA2O dataset. Having the Whois pitfalls in mind and the reference dataset by hand, our next goal is designing an automatic approach to produce an improved dataset of AS-to-organization mappings with more accurate inferences of sibling ASes.

5 Towards automatically improving inferences

In this section, we introduce our design of the automatic approach to reproduce the reference dataset. We have learned from our semi-manual investigation that matching keywords is efficient to confirm relations between organization-organization and AS-organization. Thus, we propose a clustering approach based on the keyword-matching method. In general, the approach constructs a graph for each pool, where pool elements are converted to nodes (*i.e.,* either an ASN, a CA2O.org, or a PDB.org). For each node, we collect a set of identification features and populate edges between related nodes by matching keywords. In addition, we implement different graph initialization strategies based on the knowledge of Whois pitfalls analyzed in § 4. Finally, we identify clusters of each graph and output sibling relations as well as related organizations.

We present the scope of application of the approach in § 5.1, and an overview of the approach in § 5.2. The data preparation and strategies of graph initialization are detailed in § 5.3 and § 5.4. In § 5.5 and § 5.6, we show the methods of keyword-matching and cluster discovery. Lastly, we evaluate the ability of our automatic approach on reconstructing the reference dataset in § 5.7.

5.1 Scope of application

Similar to the investigation, we only take into account disagreements between PDB and CA2O as indications of possible mistakes, thus pools without any discrepancies fall outside the scope of our approach. As shown in § 4.4, CA2O is quite accurate for pools that do not contain multiple APNIC-delegated ASes in *Class-1* (*i.e.,* a subset of non-APNIC-LIR organizations). Therefore, we simply use the CA2O mappings for these 454 pools without applying our method. As a result, we run our approach on 507 pools and 4,550 ASNs, including all pools in *Class-2* and *Class-3*, as well as the candidate APNIC-LIR pools in *Class-1*.

5.2 Method overview

Our approach consists of five stages. Initially, we build a graph in which the nodes are the ASNs, CA2O.orgs, and PDB.orgs of a pool. Then, we conduct a three-step data preparation to extract a set of identification features for each node. To complete the graph initialization, we design and implement different strategies for different classes, including pre-populating edges and sometimes adding new organization nodes. Afterwards, we examine each pair of nodes and populate edges between them if any matching keywords are found in the two sets of features. Finally, we run a Breadth-First Search algorithm on each graph and output connected components as clusters of sibling ASes and corresponding mapped organizations.

Table 4: An Overview of Collected Attributes

Attributes	ID	Name	Alias	Descr	Admin	Website
ASN	✗	✓*	✓*	✓*	✓*	✓*+
CA2O.org	✓*	✓	✗	✗	✗	✗
PDB.org	✗	✓	✓*	✗	✗	✓*

* Not always available; + from multiple sources.

5.3 Data preparation

In the following paragraphs, we introduce three data preparation steps in a sequence of data collection, data cleaning, and feature extraction. Upon completion of data preparation, we associate each node in a graph with a set of keywords.

Data collection Given that it is difficult for the automatic approach to take advantage of the Google search engine and consulting ground truths from Internet operators (as what we do in manual labeling), we partially compensate for it by leveraging more informative fields. As shown in Table 4, we collect 6 types of attributes from Whois and PDB: *ID*, *Name*, *Alias*, *Descr*, *Admin*, and *Website*, where we supplement the *website* attribute with data from BGP.tools as well.

For ASN nodes, we collect fields of *AS-name*, *descr*, and *admin-c* from Whois, where the *admin-c* field relates to the administrative contact, and the *descr* field contains auxiliary information (remind that the two fields could reveal the actual owners of customer ASes). In addition, we collect *alias* from the *AKA* (*i.e.,* also known as) field of PDB, where some Internet operators record aliases of ASes. This field might help identify relations between objects involved in acquisitions or mergers. At last, we collect *website URLs* from both PDB and BGP.tools (18,885 websites from PDB, 9,422 from BGP.tools), where ASes operated by different groups of an organization might use the same website URL.

For organization nodes, we collect *orgID* and *org-name* from Whois databases for CA2O.orgs. From PDB data, we collect *org-name*, *alias* (from *AKA* field), and *website* (13,357 websites) for PDB.orgs.

However, there are a few special cases that we should take into consideration. During the manual investigation, we notice 6 APNIC LIRs and 1 APNIC NIR register all the customer ASes with their own *admin-c*. To ensure the accuracy of our final dataset, we do not gather *admin-c* for ASes in these 7 pools (Appendix § C). Such manual input of prior information requires consistent updating.

Data cleaning First, we convert non-English characters to English characters using Python *unidecode* package to simplify the following keyword comparison. For *descr* fields with multiple lines, we adopt the same approach as CA2O and only retain the first line, as the subsequent lines are unlikely to pertain to the names of organizations (*e.g.,* street addresses, city names, etc.).

In addition, we notice that some website URLs are not up-to-date, as they automatically redirect to a different domain. Given that new domains may reveal more information, we employ *Selenium* in Python to scrape updated website URLs. For example, AS199422 records http://rezopole.net/ as its website URL in PDB, however, this URL is redirected to https://www.lyon.franceix.net/fr/, because *Rezopole* was merged to *FranceIX* in December 2020 [10]. As a result, we updated the website information of 1,880 ASes and 241 PDB.orgs.

Feature extraction We define three functions to extract features from the cleaned attributes: *SLD()* is to extract the second-level domain from *website*; *Brands()* and *Acronyms()* are to extract keywords from *all other attributes*.

SLD() We use *tldextract* Python module on the websites to extract the second-level domains. For example, we extract *franceix* as a keyword of AS199422 by using the function on the website URL https://www.lyon.franceix.net/fr/.

Brands() The function is designed to extract representative keywords, especially brand names. We use regular expressions to extract a set of English keywords containing at least two characters. Then, we filter them against manually-built lists of stop-words to eliminate words without any representative information (*e.g., llc, university, services*). For example, the output keywords of *Netflix Streaming Services Inc.* are {*netflix, streaming*}. We put the details of this function in Appendix § B.

Acronyms() Due to different conventions of naming, it is common that an AS is named by the initials of its organization name. Thus, we extract two types of possible acronyms for organization nodes: *(i)* the concatenation of upper case English characters, *(ii)* the concatenation of the first letter of each English word split by space. For example, the acronym of organization *Internet Systems Consortium, Inc.* is *isci*.

At the end of data preparation, we group all features extracted by the three functions into a keyword set and attach it with the corresponding node.

5.4 Graph initialization

We integrate two types of prior knowledge in the graph initialization stage. The first knowledge is *agreements of siblings* between CA2O and PDB: if in a pool, two ASes are recognized as siblings by both CA2O and PDB, we connect them with an edge between the ASN nodes. The second knowledge is from our investigation: as shown in the following paragraphs, we apply different strategies on pools from different classes based on the combinations of Whois pitfalls. Figure 4 illustrates one example of an initialized graph for each class.

Class-1 Our approach only applies to the pools with multiple APNIC-delegated ASes in *Class-1*, which are potentially impacted by the APNIC-LIR issue. Given that the mappings of CA2O on customer ASes are unreliable, we need to independently establish relationships between organizations and ASes. Consequently, we initialize each graph without any AS-organization links from CA2O. In addition,

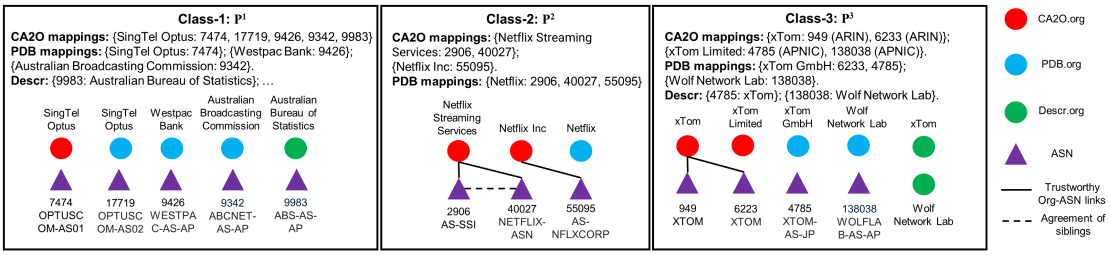

Fig. 4: Examples of graph initialization

we separate the *descr* fields from ASN nodes and initialize them as individual potential organization nodes (*i.e.,* Descr.org) according to what we learned from some APNIC LIRs in § 4.3.

Class-2 The multi-orgID issue is the only potential problem in pools of *Class-2*, where CA2O may miss some relations between organizations. During the investigation, we discovered that the existing mappings within each CA2O.org are quite reliable. Thus, we need to find potential relations between different organization nodes and merge them into bigger clusters. To this end, we keep edges between AS-organization according to the CA2O mappings. By doing so, each CA2O.org and its ASes from CA2O are connected in the initialized graph.

Class-3 Though both pitfalls might exist in pools of *Class-3*, only the CA2O.orgs with multiple APNIC-delegated ASes are possibly impacted by the APNIC-LIR issue. For these CA2O.orgs, we use the same strategy as *Class-1* to discard AS-organization links of CA2O and add *Descr.orgs* as potential organizations. For the other organizations, we use the same strategy as *Class-2* to keep the CA2O mappings.

5.5 Keyword matching

So far, we have initialized a graph with four types of nodes and some edges, where each node is associated with a keyword set. In this stage, we compare every pair of nodes (*i.e.,* ASN-ASN, ASN-Org, Org-Org) and populate an edge if there is any same keyword between the two sets. The criterion we used to compare keywords is a *keyword prefix matching*: if one word in a keyword set is equal to or is the prefix of any word in another keyword set, we consider the two nodes to be related. We do not use simple matching because it might miss some relations. For example, the *keyword prefix matching* can find relations between *Internet Systems Consortium, Inc.* and AS5277 (ISC-F-AS), since the keyword of AS5277 (isc) is the prefix of the acronym of the organization (isci). We emphasize that the risk of mismatching two randomly unrelated organizations is minimized since the Pool Detection pipeline narrows down the problem scope to related organizations according to CA2O and PDB.

Table 5: Reconstruction rate of our automatic approach, where A refers to the APNIC LIR issue and M refers to the multi-orgID issue.

	Class-1 (A)	**Class-2 (M)**	**Class-3 (A+M)**		**Manual (A)**	**Total**
Reference dataset	194	1,770	313	1,422	73	3,772
Automatic approach	194	1,649	290	1,189	73	3,395
Reconstruction rate	100%	93%	93%	84%	100%	90%

5.6 Cluster discovery

After comparing every pair of nodes by keyword matching, the final step is to identify clusters of ASes and organizations on the graph that have been created. We define connected components (CCs) as clusters, where each CC is a set of nodes that are linked to each other by paths. To find CCs, we run a Breadth-First Search algorithm on each graph. For each ASN node, the other ASN nodes in the same cluster are its siblings, and the organizations (from CA2O or PDB or Descr) in the same cluster are the inferred organizations.

5.7 Evaluation

We evaluate the performance of our automatic approach by comparing the clustering results with our manually labeled reference dataset. We use the reconstruction rate as the metric to measure the performance of our automatic approach. In Table 5, we show the numbers of mappings corrected by the dataset and approach, and the reconstruction rate for each class. The *manual* column records the corrected mappings of the 8 APNIC LIRs for which our Pool Detection finds no disagreements. In the table, we use A to refer to corrections on the APNIC-LIR issue and use M to refer to corrections on the multi-orgID issue. As a result, the sibling relations corrected by our approach involve 3,395 ASes, where the reconstruction rate is around 90% of our manual effort (3,772 ASes involved).

Our automatic approach successfully recognizes and corrects problematic CA2O mappings influenced by the APNIC-LIR issue, where we reconstruct about 96% mappings of the reference dataset. The approach fails to identify 23 customer ASes in the biggest pool from *Class-3* and incorrectly considers them as siblings. The reason is that most of the ASes are owned by individuals in China, whose names are written by Chinese PINYIN [6], which causes our keyword-matching method to mistake relations between nodes by either keywords or acronyms.

For the multi-orgID issue, our automatic approach correctly identifies *all missing siblings* for 89% mappings of the reference dataset. The approach also partially identifies missing siblings for 56 mappings in *Class-2* and 76 mappings in *Class-3*. For example, a pool in *Class-2* contains 3 CA2O.orgs (*BeeksFX VPS USA Inc.*; *Network Foundations LLC*; *Beeks Financial Cloud Ltd*), and our approach correctly recognizes 10 siblings in two *Beeks* organizations, only

[6] The official romanization system for Standard Mandarin Chinese in China

missing AS36242 (NFLLC-EQUINIX-ED) from *Network Foundations*. In fact, *Network Foundations* has another name *VDIware*, and Beeks acquired it in 2015.

The errors made by our automatic approach resulted in not identifying 354 missing siblings and mistakenly mapping 209 ASes to incorrect organizations. Out of the incorrect mappings, 193 were found in the *Class-3* pools. This is due to the fact that *Class-3* pools have a large number of ASes, and when our method incorrectly identifies the connections between organizations, it leads to a substantial amount of errors that impact a significant number of ASes. For example, a *Class-3* pool contains 3 CA2O.orgs (*DE-CIX North America Inc.*; *DE-CIX Management GmbH*; *COMNET BILGI ILETISIM TEKNOLOJILERI TICARET A.S.*), where *DE-CIX* is an organization operating Internet Exchange Points, and *COMNET* is an Internet service provider in Istanbul. Similar to the *LINX* and *STC* example, *DE-CIX* operates AS47298 (ISTIX) together with *COMNET*. Since the features of AS47298 contain keywords related to both organizations, our approach mistakes all 30 ASes in this pool as siblings.

6 Case Study: MOAS Event Analysis

In this section, we present a case study of BGP hijacking analysis to illustrate the relevance of our sibling dataset. We focus on Multiple Origin AS (MOAS) events, which are potentially linked to one type of BGP hijacking attack. A MOAS event occurs when in BGP, an IP prefix appears to be originated from more than one ASes [27]. In this context, sibling relationships between involved ASes provide key information to understand the event, the likelihood of misconfiguration and to eventually start a forensic investigation. For instance, the sibling relationship between involved ASes is an important factor when determining if an event is malicious or not. If no other suspicious behaviors are detected (*e.g.,* the AS is infiltrated by attackers), the events between sibling ASNs are highly possible to be non-malicious. As a case study, we collect all 97,975 MOAS events (containing 30,709 pairs of ASNs) monitored by the Global Routing Intelligence Platform [5] in 2021 and compare the results of using our dataset or CA2O on identifying events that happened between sibling ASes.

Using our dataset we discover more sibling-related events, also identify several non-sibling related events which CA2O identifies as sibling-related events. Both our dataset and CA2O agree on 2,076 pairs of ASes being siblings. However, our dataset additionally identifies 17% more pairs of sibling ASNs, with a total of 360 pairs and 4,219 events. We list some examples in Table 6, where the sibling relationship discovered by our dataset provides more context to the events. In addition, using our method, we identify 11 MOAS events that happened between ASes of APNIC Local Internet Registries (LIRs) and customer ASes, which CA2O considers as sibling-related events. Our dataset provides a more precise interpretation for these events: it is possible that the LIR serves as the upstream and originates the prefix in BGP for its customer. For example, our dataset identifies an event that happened between AS9658 and AS131212, where the former AS belongs to an LIR (*Eastern Telecommunications Philippines*), and the

Table 6: Examples of newfound sibling ASes in MOAS events

AS-pair	CA2O.orgs	#Occurrence
3356, 3561	Level 3 Parent LLC; CenturyLink Communications LLC	4
4755, 6453	Tata Communications Limited; TATA COMMUNICATIONS (AMERICA) Ltd	6
20115, 20001	Charter Communications; Charter Communications Inc	3
19527, 139190	Google LLC; Google Asia Pacific Pte. Ltd	3
36617, 10515	VeriSign Global Registry Services; VeriSign Infrastructure & Operations	75
16625, 20940	Akamai Technologies Inc; Akamai International B.V.	28
33438, 12989	Highwinds Network Group Inc; StackPath LLC	14
8190, 3257	GTT; GTT Communications Inc	1

latter belongs to a customer AS (*Robinsons Land Corporation*). To conclude, our output dataset contributes to a more accurate understanding of BGP hijacking events and better supports potential forensic investigations.

7 Discussion and future work

7.1 Limitations of our methodology

There are some inherent limitations caused by the dependency on PDB. First of all, PDB is not specialized for the AS-to-organization mapping similar to Whois, hence there is no guarantee on the detected pools to perfectly follow our definition. For example, operators of some subsidiaries might consider their business independent, so they maintain different organizations in PDB just like the information in the Whois databases. As a result, our Pool Detection pipeline is not able to discover the relations between the involved ASes. For example, *Singtel Optus Pty Limited* (*Optus*) and *Singapore Telecommunications Limited* (*Singtel*) are two individual organizations according to both PDB and CA2O, even though *Optus* is a completely owned subsidiary of *Singtel*. As a result, our Pool Detection places them into two different pools. Another example is *vodafone*, which owns and operates networks all over the world. There are 35 and 29 different organizations containing *vodafone* as the brand name in CA2O and PDB respectively, where our Pool Detection only recognizes a few subsets of them in the same pool (*e.g., Vodafone UK Limited Mobile AS* and *Vodafone UK Limited*) but neglects most of them. One possible solution to the limitation is implementing the keyword matching method on all ASes and organizations to find possible relations. However, more careful validation is needed because organizations without any relation (especially in different countries) may have exactly the same brand names.

Another limitation is that the information hidden in the natural language data of PDB (*e.g.,* the *notes* field) is hard to extract. For example, the *notes* of AS137945 in PDB records the following information: "*Nutrien operates AS137945 in APAC and AS393891 in North America; AS137900 is also operated by Nutrien APAC.*" The information about AS137900 is accurate. Indeed, AS137900 is mapped to *Ruralco Holdings Limited* by Whois, where *Nutrien* acquired *Ruralco*

in 2020. However, since AS137900 is not registered in PDB, and it does not have any sibling according to CA2O, our Pool Detection isolates AS137900 from the other two ASes of *Nutrien*. If the relations between ASes and organizations could be correctly extracted from the natural language data, the Pool Detection could become more precise as well as the automatic clustering approach. Towards improving this problem, leveraging natural language processing methods might be one possible solution.

Even though the knowledge in PDB is fully leveraged, some information is still not covered by the datasets we used, especially for mergers and acquisitions. There are some commercial databases such as *Crunchbase* and *Dun & Bradstreet* which contain plenty of information such as acquisition history, and subsidiary list. As for the drawbacks, the databases are neither authoritative nor directly maintained by the operators, and it is hard to validate the information.

7.2 Interaction with Internet operators

Given that the ground truth of AS-to-organization mappings can be only obtained from Internet operators, a virtuous interaction with Internet operators is extremely beneficial. When constructing the reference dataset, we contacted 105 Internet operators. Except for 10 undeliverable email addresses, we received 12 replies in total. On the one hand, RIRs need to impose more precise supervision to ensure operators update the contact information as soon as the email changes. On the other hand, our researchers need to do more work to facilitate active and constructive interactions with Internet operators. We plan to create a website for our project about the mappings that we found different from the Whois records and welcome the authorized operators (*i.e.,* with a PDB account) to verify or modify the data, which could also help us to update our dataset and approach. Another aspect of interaction is encouraging the operators to maintain and update the information in user-maintained public databases like PDB and BGP.tools, which are extremely helpful for researchers (remind that only around 23% ASes are currently registered in the PDB database). Our approach could benefit from the higher AS coverage of PDB to attain more complete results of Pool Detection and collect more features for the automatic approach. An open question is how to motivate Internet operators to maintain the databases.

7.3 Extension the mappings for AS-level analysis

Our definitions of ownership and sibling ASes mainly aim for applications related to Internet behaviors at an organizational level, which may not fit AS-level studies perfectly. For example, although *CenturyLink* acquired *Level 3* and then renamed to *Lumen* in 2020, business types between the two divisions are quite different: ASes (previously) operated by *Level 3* are mainly for transit purposes (*e.g.,* AS3356), while ASes (previously) operated by *CenturyLink* are mainly for residential Internet services (*e.g.,* AS3561). In this scenario, separating

ASes of these two divisions could benefit the AS-level analysis, such as AS-type classification. One possible solution is a hierarchical structure of AS-to-organization mappings which also takes the subsidiaries and divisions into account. A preliminary but not verified structure is to organize a tree-like hierarchy for the pools impacted by the multi-orgID issue, where we place our reference organization(s) at the top level, CA2O.orgs at the middle level, and ASes at the bottom as leaves. Consequently, the AS-level analysis only focuses on the middle and bottom layers, while information on sibling relations between the subsidiaries is maintained at the top layers. We leave the evaluation of the necessity and effects of such hierarchical mappings as future work.

8 Conclusions

In this work, we aim to improve the inferences of sibling relations in AS-to-organization mappings to benefit Internet researchers. We start by comparing the state-of-the-art dataset CAIDA AS-to-organization with Whois databases and show that CA2O is susceptible to wrong inferences due to inaccurate information in Whois databases. Then we leverage PeeringDB data to find the potentially problematic mappings and conduct a meticulous semi-manual investigation. During the process, we identify two pitfalls in Whois: the APNIC-LIR issue and the multi-orgID issue, which are the main causes of the inaccuracies. We also construct a reference dataset that corrects 12.5% CA2O organizations that have sibling ASes. We further propose an automatic and scalable approach to reproduce the dataset with high fidelity, which is able to automatically improve inferences of sibling ASes for each new version of CA2O.

9 Ethics

This work does not raise any ethical issues.

References

1. Bgp.tools. https://bgp.tools/
2. The CAIDA AS Organizations Dataset, (downloaded on July 1, 2022). http://www.caida.org/data/as-organizations
3. CAIDA AS Rank. http://as-rank.caida.org/
4. Daily snapshots of PeeringDB data, (downloaded on April 4, 2022). https://publicdata.caida.org/datasets/peeringdb/
5. Global Routing Intelligence Platform (GRIP). http://grip.inetintel.cc.gatech.edu/
6. GTT acquired Interoute in 2018. https://www.gtt.net/us-en/media-center/press-releases/gtt-to-acquire-interoute/
7. The Internet registry system. www.ripe.net/participate/internet-governance/internet-technical-community/the-rir-system
8. Introduction to ARIN's databases. https://www.arin.net/resources/guide/account/database/

9. Mapping Autonomous Systems to Organizations: CAIDA's Inference Methodology. https://www.caida.org/archive/as2org/
10. The merge of France IX and Rezopole A.D. https://www.linkedin.com/company/france-ix
11. The merger of Nutrien in 2018. https://www.nutrien.com/investors/news-releases/2018-agrium-and-potashcorp-merger-completed-forming-nutrien-leader-global/
12. News of the Acquisition of CNS by Beeks. https://beeksgroup.com/news/beeks-acquires-vps-provider-cns/
13. News of the Acquisition of Linode by Akamai. https://www.akamai.com/newsroom/press-release/akamai-completes-acquisition-of-linode
14. Process of ASN application of RIPE. https://www.ripe.net/manage-ips-and-asns/as-numbers/request-an-as-number
15. The public datasets of ISI ANT lab. https://ant.isi.edu/datasets/all.html
16. Template of APNIC Whois. www.apnic.net/manage-ip/using-whois/guide/aut-num/
17. Cai, X., Heidemann, J., Krishnamurthy, B., Willinger, W.: Towards an AS-to-organization Map. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. pp. 199–205 (2010)
18. Cai, X., Heidemann, J., Krishnamurthy, B., Willinger, W.: An Organization-Level View of the Internet and its Implications(extended) p. 26 (Jun 2012)
19. Dainotti, A., Benson, K., King, A., Huffaker, B., Glatz, E., Dimitropoulos, X., Richter, P., Finamore, A., Snoeren, A.C.: Lost in space: improving inference of ipv4 address space utilization. IEEE Journal on Selected Areas in Communications **34**(6), 1862–1876 (2016)
20. Jin, Y., Scott, C., Dhamdhere, A., Giotsas, V., Krishnamurthy, A., Shenker, S.: Stable and Practical {AS} Relationship Inference with ProbLink. pp. 581–598 (2019), https://www.usenix.org/conference/nsdi19/presentation/jin
21. Konte, M., Perdisci, R., Feamster, N.: ASwatch: An AS Reputation System to Expose Bulletproof Hosting ASes (2015)
22. Liu, J., Yang, B., Liu, J., Lu, Y., Zhu, K.: A Method of Route Leak Anomaly Detection Based on Heuristic Rules. pp. 662–666. Atlantis Press (Jun 2017). https://doi.org/10.2991/ammee-17.2017.127, https://www.atlantis-press.com/proceedings/ammee-17/25878482, iSSN: 2352-5401
23. Luckie, M., Huffaker, B., Dhamdhere, A., Giotsas, V., claffy, k.: AS relationships, customer cones, and validation. In: Proceedings of the 2013 conference on Internet measurement conference - IMC '13. pp. 243–256. ACM Press, Barcelona, Spain (2013). https://doi.org/10.1145/2504730.2504735, http://dl.acm.org/citation.cfm?doid=2504730.2504735
24. Nemmi, E.N., Sassi, F., La Morgia, M., Testart, C., Mei, A., Dainotti, A.: The parallel lives of Autonomous Systems: ASN allocations vs. BGP. In: Proceedings of the 21st ACM Internet Measurement Conference. pp. 593–611. IMC '21, Association for Computing Machinery, New York, NY, USA (Nov 2021). https://doi.org/10.1145/3487552.3487838, https://doi.org/10.1145/3487552.3487838
25. Padmanabhan, R., Filastò, A., Xynou, M., Raman, R.S., Middleton, K., Zhang, M., Madory, D., Roberts, M., Dainotti, A.: 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. pp. 27–36 (2021)
26. Testart, C., Richter, P., King, A., Dainotti, A., Clark, D.: Profiling BGP Serial Hijackers: Capturing Persistent Misbehavior in the Global Routing Table. In: Proceedings of the Internet Measurement Conference on - IMC '19. pp. 420–434.

ACM Press, Amsterdam, Netherlands (2019). https://doi.org/10.1145/3355369.3355581, https://dl.acm.org/doi/10.1145/3355369.3355581
27. Zhao, X., Pei, D., Wang, L., Massey, D., Mankin, A., Wu, S.F., Zhang, L.: An analysis of BGP multiple origin AS (MOAS) conflicts. In: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement. pp. 31–35 (2001)
28. Ziv, M., Izhikevich, L., Ruth, K., Izhikevich, K., Durumeric, Z.: ASdb: a system for classifying owners of autonomous systems. In: Proceedings of the 21st ACM Internet Measurement Conference. pp. 703–719. ACM, Virtual Event (Nov 2021). https://doi.org/10.1145/3487552.3487853, https://dl.acm.org/doi/10.1145/3487552.3487853

A Information of RIR/NIR Whois

APNIC The bulk Whois data of APNIC is public, while among 7 NIRs, only JPNIC and KRNIC publish their bulk Whois data. We learned from the APNIC helpdesk that if NIRs make further assignments within the NIR-maintained whois database, they may not be reflected in the APNIC Whois database.

20,127 ASes are delegated in the APNIC region including the ones delegated by the NIRs. *aut-num* (*i.e.,* autonomous system number) and *organisation* are the AS-object and org-object in APNIC Whois, associated with the *org* field (*i.e.,* org-id of the organization) in *aut-num*. However, 8,781 ASes in APNIC do not have *org* field (*i.e.,* no related organization objects), where 99.4% of such ASes are registered in the countries of 7 NIRs. For these ASes, the *descr* (*i.e.,* description) field in AS-objects carries the name of the owner organization without association by org-id. The *descr* field is mandatory [16], and all AS-objects have such field including the ones with associated organization-objects.

RIPE NCC The bulk Whois data of RIPE is public. 37,672 ASes are delegated in the RIPE region, which is the most among the 5 RIRs. RIPE NCC has a similar structure as APNIC that there are *aut-num* and *organization* objects associated by org-id. Though no NIR exists in the RIPE region, there is still a small amount of ASes (108 ASNs) without associated organizations, whose holder organization is in the *descr* field. Different from APNIC, the *descr* field is not mandatory and only 3,962 ASes in RIPE have this field.

AFRINIC The bulk Whois data of AFRINIC is public. AFRINIC allocates the least AS numbers among RIRs, where only 2,168 ASes are delegated in the AFRINIC region. The Whois structure of AFRINIC is similar to APNIC and RIPE but more consistent: all *aut-num* objects have *org* fields associating with org-objects and the *descr* field is also mandatory in AFRINIC.

ARIN The access to ARIN bulk Whois data needs an application (we get access for this work). 31,446 ASes are delegated in the ARIN region. ARIN uses its own format of Whois [8]: *ASHandle* and *OrgName* are two main objects, associated by *OrgID*. AS-objects does not have the *descr* field and every ASN-object has an associated org-object.

LACNIC The access to LACNIC bulk Whois data needs an application (we do not get access for this work). 12,740 ASes are delegated in the LACNIC region. To compare CA2O with LACNIC Whois, we conduct a web scraping on the LACNIC official webpage for Whois to collect the Whois mappings.

B Details of Keywords function

We implement two lists of stop-words, where the first list contains the words that can not be used to identify an organization, while the second list might be useful for some time. The first list contains *apnic, enterprise, asn, sas, as, information, ap, pvt, university, jpnic, jsco, telecom, and, bvba, autonomous, ltda, services, for, op, backbone, telekom, based, ohg, de, gmbh, technologies, lacnic, pt, legacy, inc, company, the, technology, of, llc, sdn, organization, afrinic, com, idnic, bhd, da, international, corporation, twnic, limited, research, or, aka, pty, service, solutions, me, arin, ltd, jsc, in, org, ripe.*

The second list contains *health, communication, tecnologia, data, network, comunicacao, center, coop, hospital, australia, bank, servi, servers, sg, telecomunica, el, northern, north, net, en, me, systems, sdn, telecommunications, telecomunices, telecommunication, east, eu, uab, education, info, de, public, silva, exchange, world, serv, college, communications, eng, western, digital, hosting, apac, city, southern, yue, internet, broadband, asia, link, route, uk, consumo, provedora, networks, japan, tech, ag, west, sp, cloud, web, co, telecomunicacoes, os, servicos, ab, ix, comunica, tel, publicos, telefon, experimental, yu, europe, connect, eastern, south, computing, group, county, global.* In addition, we add the names of countries and the two-letter country codes to the second list.

For each set of extracted English keywords, we first filter out the words in the first list. Then we examine if all the remaining words exist in the second list. If so, we do not use the second list; otherwise, we use the list to filter out part of the words.

C Manual input knowledge

C.1 Manual input pools in § 4

We identified 8 CA2O.orgs during the semi-manual investigation, which are likely to be APNIC LIRs (211 ASes involved). The pool detection did not recognize them because none of the involved ASes maintain information in PDB. We list the names of them here: *REANNZ Education and Schools*; *Internet Thailand Company Ltd.*; *ePLDT Inc.*; *CS Loxinfo Public Company Limited*; *Globe Telecom (GMCR,INC)*; *Sky Internet*; *KSC Commercial Internet Co.Ltd.*; *Philippine Long Distance Telephone Co.*

C.2 Manual knowledge of *admin-c* in § 5

We identified several pools that the CA2O.orgs are very likely to be APNIC LIRs, but the involved ASes have the same *admin-c* fields. For the sake of the accuracy of our dataset, we do not add *admin-c* as a feature for the ASes in these pools:

One pool containing of an NIR. IRINN (Indian Registry for Internet Names and Numbers) put their org-handle (RB486-AP) in *admin-c* fields for 11 ASes. We contacted IRINN and confirmed that it was a technical glitch that the system automatically set the IRINN nic-handle on the ASes delegated by IRINN if Whois server issue happened.

Six pools containing APNIC LIRs. We list the names of the APNIC LIRs here: *United Information Highway*; *Eastern Telecommunications Philippines, Inc.*; *SingTel Optus Pty Ltd*; *True Internet Co.,Ltd. and TRUE INTERNET*; *Communications & Communicate Nepal Pvt Ltd*; *VOCUS PTY LTD.*